# SEQUENCE-DETERMINED DNA FRAGMENTS AND CORRESPONDING POLYPEPTIDES ENCODED THEREBY

09/691,635

This application claims priority under 35 USC §119(e), §119(a-d) and §120 of the following applications, the entire contents of which are hereby incorporated by reference:

| Country | Filing Date | Attorney No. | Client No. | Application No. |
|---|---|---|---|---|
| United States | 09/28/1999 | 2750-0563P | 00106.001 | 60/156,458 |

This application contains a CDR, the entire contents of which are hereby incorporated by reference. The CDR contains the following files:

| File Name | Date of Creation | File Size |
|---|---|---|
| 2750-1026P.ST25 | May 31, 2001 | 106 KB |

## FIELD OF THE INVENTION

The present invention relates to isolated polynucleotides that represent a complete gene, or a fragment thereof, that is expressed. In addition, the present invention relates to the polypeptide or protein corresponding to the coding sequence of these polynucleotides. The present invention also relates to isolated polynucleotides that represent regulatory regions of genes. The present invention also relates to isolated polynucleotides that represent untranslated regions of genes. The present invention further relates to the use of these isolated polynucleotides and polypeptides and proteins.

## DESCRIPTION OF THE RELATED ART

Efforts to map and sequence the genome of a number of organisms are in progress; a few complete genome sequences, for example those of *E. coli* and *Saccharomyces cerevisiae* are known (Blattner et al., *Science* 277:1453 (1997); Goffeau et al., *Science* 274:546 (1996)). The complete genome of a multicellular organism, *C. elegans*, has also been sequenced (See, the *C. elegans* Sequencing Consortium, *Science* 282:2012 (1998)). To date, no complete genome of a plant has been sequenced, nor has a complete cDNA complement of any plant been sequenced.

## SUMMARY OF THE INVENTION

The present invention comprises polynucleotides, such as complete cDNA sequences and/or sequences of genomic DNA encompassing complete genes, fragments of genes, and/or regulatory elements of genes and/or regions with other functions and/or intergenic regions, hereinafter collectively referred to as Sequence-Determined DNA Fragments (SDFs), from

different plant species, particularly corn, wheat, soybean, rice and *Arabidopsis thaliana,* and other plants and or mutants, variants, fragments or fusions of said SDFs and polypeptides or proteins derived therefrom. In some instances, the SDFs span the entirety of a protein-coding segment. In some instances, the entirety of an mRNA is represented. Other objects of the invention that are also represented by SDFs of the invention are control sequences, such as, but not limited to, promoters. Complements of any sequence of the invention are also considered part of the invention.

Other objects of the invention are polynucleotides comprising exon sequences, polynucleotides comprising intron sequences, polynucleotides comprising introns together with exons, intron/exon junction sequences, 5' untranslated sequences, and 3' untranslated sequences of the SDFs of the present invention. Polynucleotides representing the joinder of any exons described herein, in any arrangement, for example, to produce a sequence encoding any desirable amino acid sequence are within the scope of the invention.

The present invention also resides in probes useful for isolating and identifying nucleic acids that hybridize to an SDF of the invention. The probes can be of any length, but more typically are 12-2000 nucleotides in length; more typically, 15 to 200 nucleotides long; even more typically, 18 to 100 nucleotides long.

Yet another object of the invention is a method of isolating and/or identifying nucleic acids using the following steps:

(a) contacting a probe of the instant invention with a polynucleotide sample under conditions that permit hybridization and formation of a polynucleotide duplex; and

(b) detecting and/or isolating the duplex of step (a).

The conditions for hybridization can be from low to moderate to high stringency conditions. The sample can include a polynucleotide having a sequence unique in a plant genome. Probes and methods of the invention are useful, for example, without limitation, for mapping of genetic traits and/or for positional cloning of a desired fragment of genomic DNA.

Probes and methods of the invention can also be used for detecting alternatively spliced messages within a species. Probes and methods of the invention can further be used to detect or isolate related genes in other plant species using genomic DNA (gDNA) and/or cDNA libraries. In some instances, especially when longer probes and low to moderate stringency hybridization conditions are used, the probe will hybridize to a plurality of cDNA and/or gDNA sequences of a plant. This approach is useful for isolating representatives of gene families which are identifiable by possession of a common functional domain in the gene product or which have

common cis-acting regulatory sequences. This approach is also useful for identifying orthologous genes from other organisms.

The present invention also resides in constructs for modulating the expression of the genes comprised of all or a fragment of an SDF. The constructs comprise all or a fragment of the expressed SDF, or of a complementary sequence. Examples of constructs include ribozymes comprising RNA encoded by an SDF or by a sequence complementary thereto, antisense constructs, constructs comprising coding regions or parts thereof, constructs comprising promoters, introns, untranslated regions, scaffold attachment regions, methylating regions, enhancing or reducing regions, DNA and chromatin conformation modifying sequences, etc. Such constructs can be constructed using viral, plasmid, bacterial artificial chromosomes (BACs), plasmid artificial chromosomes (PACs), autonomous plant plasmids, plant artificial chromosomes or other types of vectors and exist in the plant as autonomous replicating sequences or as DNA integrated into the genome. When inserted into a host cell the construct is, preferably, functionally integrated with, or operatively linked to, a heterologous polynucleotide. For instance, a coding region from an SDF might be operably linked to a promoter that is functional in a plant.

The present invention also resides in host cells, including bacterial or yeast cells or plant cells, and plants that harbor constructs such as described above. Another aspect of the invention relates to methods for modulating expression of specific genes in plants by expression of the coding sequence of the constructs, by regulation of expression of one or more endogenous genes in a plant or by suppression of expression of the polynucleotides of the invention in a plant. Methods of modulation of gene expression include without limitation (1) inserting into a host cell additional copies of a polynucleotide comprising a coding sequence; (2) modulating an endogenous promoter in a host cell; (3) inserting antisense or ribozyme constructs into a host cell and (4) inserting into a host cell a polynucleotide comprising a sequence encoding a variant, fragment, or fusion of the native polypeptides of the instant invention.

**BRIEF DESCRIPTION OF THE TABLES**

In TABLE 1, the format of the data is as follows:

In Table 1, sequence data are presented in the form of annotation of a reference sequence. The format is shown below. The reference sequence is shown at the top of the annotation file as a 7 digit sequence number preceded by ">" (*e.g.* >5019261). The sequence identifier is a "gi" number that identifies a specific DNA sequence in the publically accessible BLAST Databases on the NCBI FTP web site (accessible at ncbi.nlm.gov/blast).

In particular, the "nt.Z" nucleotide sequence data base at the NCBI FTP site utilizes the "gi" identifiers to assign by NCBI a unique identifier for each sequence in the databases, thereby providing a non-redundant database for sequences from various data bases, including GenBank, EMBL, DDBJ (DNA Database of Japan) and PDB (Brookhaven Protein Data Bank). Thus, the line in TABLE 1 beginning with sequence number identifies the unique "gi" identifier followed by the corresponding GenBank (gb) accession number and locus. The reference sequence number is followed on the next line by data regarding the length of the sequence ("len") and the number of exons found in the sequence by the analysis program ("nex").

The annotation data are presented in columns; the leftmost column identifies the position of the putative exon in the gene as initial ("init"), internal ("intr") or terminal ("term"). Genes considered composed of a single exon are denoted "sngl". The next column describes the position in the nucleotide sequence beginning the exon ("start") and the next column describes the position in the nucleotide sequence ending the exon ("stop"). The direction of the gene is indicated in the next column, "+" indicating 5' - 3' in the direction presented in the database, "-" indicating the opposite orientation. The "gene number" is given in the final column. Exons having the same gene number are grouped in the order shown to create the relevant coding sequence.

>5019261 ⇐ This is the gi number of the public sequence

len = 97208 nex = 121

⇑ ⇑

Length Number exons of public sequence

| Exon Type ⇓ | Start ⇓ | Stop ⇓ | Direction ⇓ | Gene Number ⇓ |
|---|---|---|---|---|
| Sngl | 602 | 778 | + | 0 |
| Sngl | 990 | 1316 | + | 1 |
| Sngl | 2356 | 2691 | + | 2 |
| Sngl | 4634 | 4735 | + | 3 |
| Sngl | 4973 | 5092 | + | 4 |
| Sngl | 5746 | 5874 | + | 5 |
| Init | 8119 | 8798 | + | 6 |
| Term | 9284 | 9518 | + | 6 |
| Init | 10827 | 11150 | + | 7 |
| Term | 11294 | 11335 | + | 7 |
| Sngl | 12655 | 12825 | + | 8 |
| Sngl | 13303 | 13596 | + | 9 |
| Sngl | 18654 | 18782 | + | 10 |
| Sngl | 19880 | 20086 | + | 11 |
| Init | 21476 | 21539 | + | 12 |
| Intr | 21647 | 21802 | + | 12 |
| Term | 23488 | 23567 | + | 12 |
| Init | 25035 | 25133 | + | 13 |
| Intr | 25466 | 25589 | + | 13 |
| Intr | 25677 | 25786 | + | 13 |
| Intr | 25899 | 25962 | + | 13 |
| Intr | 26045 | 26109 | + | 13 |
| Intr | 26188 | 26253 | + | 13 |
| Term | 26350 | 26448 | + | 13 |
| Sngl | 27671 | 27793 | + | 14 |
| Sngl | 29126 | 29299 | + | 15 |
| Sngl | 30266 | 30364 | + | 16 |
| Sngl | 31717 | 31929 | + | 17 |
| Sngl | 32102 | 32209 | + | 18 |
| Sngl | 32450 | 32548 | + | 19 |

| | | | | | |
|---|---|---|---|---|---|
| | Sngl | 32634 | 32726 | + | 20 |
| | Init | 35603 | 35743 | + | 21 |
| | Term | 35829 | 36185 | + | 21 |
| | Init | 36954 | 37098 | + | 22 |
| 5 | Term | 38100 | 38158 | + | 22 |
| | Init | 39635 | 39944 | + | 23 |
| | Intr | 40242 | 40372 | + | 23 |
| | Intr | 40462 | 40695 | + | 23 |
| | Intr | 40815 | 41070 | + | 23 |
| 10 | Intr | 41176 | 41255 | + | 23 |
| | Intr | 42212 | 42419 | + | 23 |
| | Intr | 42940 | 43070 | + | 23 |
| | Intr | 43177 | 43410 | + | 23 |
| | Intr | 43580 | 43835 | + | 23 |
| 15 | Intr | 46672 | 46715 | + | 23 |
| | Intr | 48334 | 48532 | + | 23 |

## DETAILED DESCRIPTION OF THE INVENTION

The invention relates to (I) polynucleotides and methods of use thereof, such as

IA. Probes, Primers and Substrates;

20 IB. Methods of Detection and Isolation;

B.1. Hybridization;

B.2. Methods of Mapping;

B.3. Southern Blotting;

B.4. Isolating cDNA from Related Organisms;

25 B.5. Isolating and/or Identifying Orthologous Genes

IC. Methods of Inhibiting Gene Expression

C.1. Antisense

C.2. Ribozyme Constructs;

C.3. Chimeraplasts;

30 C.4 Co-Suppression;

C.5. Transcriptional Silencing

C.6. Other Methods to Inhibit Gene Expression

ID. Methods of Functional Analysis;

IE. Promoter Sequences and Their Use;

35 IF. UTRs and/or Intron Sequences and Their Use; and

IG. Coding Sequences and Their Use.

The invention also relates to (II) polypeptides and proteins and methods of use thereof, such as IIA. Native Polypeptides and Proteins

A.1 Antibodies

A.2 In Vitro Applications

IIB. Polypeptide Variants, Fragments and Fusions

B.1 Variants

B.2 Fragments

B.3 Fusions

The invention also includes (III) methods of modulating polypeptide production, such as

IIIA. Suppression

A.1 Antisense

A.2 Ribozymes

A.3 Co-suppression

A.4 Insertion of Sequences into the Gene to be Modulated

A.5 Promoter Modulation

A.6 Expression of Genes containing Dominant-Negative Mutations

IIIB. Enhanced Expression

B.1 Insertion of an Exogenous Gene

B.2 Promoter Modulation

The invention further concerns (IV) gene constructs and vector construction, such as

IVA. Coding Sequences

IVB. Promoters

IVC. Signal Peptides

The invention still further relates to

V Transformation Techniques

**Definitions**

Allelic variant An "allelic variant" is an alternative form of the same SDF, which resides at the same chromosomal locus in the organism. Allelic variations can occur in any portion of the gene sequence, including regulatory regions. Allelic variants can arise by normal genetic variation in a population. Allelic variants can also be produced by genetic engineering methods. An allelic variant can be one that is found in a naturally occurring plant, including a cultivar or ecotype. An allelic variant may or may not give rise to a phenotypic change, and may or may not be expressed. An allele can result in a detectable change in the phenotype of the trait represented by the locus. A phenotypically silent allele can give rise to a product.

Alternatively spliced messages Within the context of the current invention, "alternatively spliced messages" refers to mature mRNAs originating from a single gene with variations in the number and/or identity of exons, introns and/or intron-exon junctions.

Chimeric The term "chimeric" is used to describe genes, as defined supra, or contructs wherein at least two of the elements of the gene or construct, such as the promoter and the coding sequence and/or other regulatory sequences and/or filler sequences and/or complements thereof, are heterologous to each other.

Constitutive Promoter: Promoters referred to herein as "constitutive promoters" actively promote transcription under most, but not necessarily all, environmental conditions and states of development or cell differentiation. Examples of constitutive promoters include the cauliflower mosaic virus (CaMV) 35S transcript initiation region and the 1' or 2' promoter derived from T-DNA of *Agrobacterium tumefaciens,* and other transcription initiation regions from various plant genes, such as the maize ubiquitin-1 promoter, known to those of skill.

Coordinately Expressed: The term "coordinately expressed," as used in the current invention, refers to genes that are expressed at the same or a similar time and/or stage and/or under the same or similar environmental conditions.

Domain: Domains are fingerprints or signatures that can be used to characterize protein families and/or parts of proteins. Such fingerprints or signatures can comprise conserved (1) primary sequence, (2) secondary structure, and/or (3) three-dimensional conformation. Generally, each domain has been associated with either a family of proteins or motifs. Typically, these families and/or motifs have been correlated with specific *in-vitro* and/or *in-vivo* activities. A domain can be any length, including the entirety of the sequence of a protein. Detailed descriptions of the domains, associated families and motifs, and correlated activities of the polypeptides of the instant invention are described below. Usually, the polypeptides with designated domain(s) can exhibit at least one activity that is exhibited by any polypeptide that comprises the same domain(s).

Endogenous The term "endogenous," within the context of the current invention refers to any polynucleotide, polypeptide or protein sequence which is a natural part of a cell or organisms regenerated from said cell.

Exogenous "Exogenous," as referred to within, is any polynucleotide, polypeptide or protein sequence, whether chimeric or not, that is initially or subsequently introduced into the genome of an individual host cell or the organism regenerated from said host cell by any means other than by a sexual cross. Examples of means by which this can be accomplished are described below, and include *Agrobacterium*-mediated transformation (of dicots - *e.g.* Salomon et al. *EMBO J.* 3:141 (1984); Herrera-Estrella et al. *EMBO J.* 2:987 (1983); of monocots, representative papers are those by Escudero et al., *Plant J.* 10:355 (1996), Ishida et al., *Nature Biotechnology* 14:745 (1996), May et al., *Bio/Technology* 13:486 (1995)), biolistic methods (Armaleo et al., *Current Genetics* 17:97 1990)), electroporation, *in planta* techniques, and the like. Such a plant containing the exogenous nucleic acid is referred to here as a $T_0$ for the primary transgenic plant and $T_1$ for the first generation. The term "exogenous" as used herein is also intended to encompass inserting a naturally found element into a non-naturally found location.

Filler sequence: As used herein, "filler sequence" refers to any nucleotide sequence that is inserted into DNA construct to evoke a particular spacing between particular components such as a promoter and a coding region and may provide an additional attribute such as a restriction enzyme site.

Gene: The term "gene," as used in the context of the current invention, encompasses all regulatory and coding sequence contiguously associated with a single hereditary unit with a genetic function (see SCHEMATIC 1). Genes can include non-coding sequences that modulate the genetic function that include, but are not limited to, those that specify polyadenylation, transcriptional regulation, DNA conformation, chromatin conformation, extent and position of base methylation and binding sites of proteins that control all of these. Genes comprised of "exons" (coding sequences), which may be interrupted by "introns" (non-coding sequences), encode proteins. A gene's genetic function may require only RNA expression or protein production, or may only require binding of proteins and/or nucleic acids without associated expression. In certain cases, genes adjacent to one another may share sequence in such a way that one gene will overlap the other. A gene can be found within the genome of an organism, artificial chromosome, plasmid, vector, etc., or as a separate isolated entity.

Gene Family: "Gene family" is used in the current invention to describe a group of functionally related genes, each of which encodes a separate protein.

Heterologous sequences: "Heterologous sequences" are those that are not operatively linked or are not contiguous to each other in nature. For example, a promoter from corn is considered heterologous to an *Arabidopsis* coding region sequence. Also, a promoter from a gene encoding a growth factor from corn is considered heterologous to a sequence encoding the corn receptor for the growth factor. Regulatory element sequences, such as UTRs or 3' end termination sequences that do not originate in nature from the same gene as the coding sequence originates from, are considered heterologous to said coding sequence. Elements operatively linked in nature and contiguous to each other are not heterologous to each other. On the other hand, these same elements remain operatively linked but become heterologous if other filler sequence is placed between them. Thus, the promoter and coding sequences of a corn gene expressing an amino acid transporter are not heterologous to each other, but the promoter and coding sequence of a corn gene operatively linked in a novel manner are heterologous.

Homologous gene In the current invention, "homologous gene" refers to a gene that shares sequence similarity with the gene of interest. This similarity may be in only a fragment of the sequence and often represents a functional domain such as, examples including without

limitation a DNA binding domain, a domain with tyrosine kinase activity, or the like. The functional activities of homologous genes are not necessarily the same.

Inducible Promoter An "inducible promoter" in the context of the current invention refers to a promoter which is regulated under certain conditions, such as light, chemical concentration, protein concentration, conditions in an organism, cell, or organelle, etc. A typical example of an inducible promoter, which can be utilized with the polynucleotides of the present invention, is PARSK1, the promoter from the *Arabidopsis* gene encoding a serine-threonine kinase enzyme, and which promoter is induced by dehydration, abscissic acid and sodium chloride (Wang and Goodman, *Plant J.* 8:37 (1995)) Examples of environmental conditions that may affect transcription by inducible promoters include anaerobic conditions, elevated temperature, or the presence of light.

Intergenic region "Intergenic region," as used in the current invention, refers to nucleotide sequence occurring in the genome that separates adjacent genes.

Mutant gene In the current invention, "mutant" refers to a heritable change in DNA sequence at a specific location. Mutants of the current invention may or may not have an associated identifiable function when the mutant gene is transcribed.

Orthologous Gene In the current invention "orthologous gene" refers to a second gene that encodes a gene product that performs a similar function as the product of a first gene. The orthologous gene may also have a degree of sequence similarity to the first gene. The orthologous gene may encode a polypeptide that exhibits a degree of sequence similarity to a polypeptide corresponding to a first gene. The sequence similarity can be found within a functional domain or along the entire length of the coding sequence of the genes and/or their corresponding polypeptides.

Percentage of sequence identity "Percentage of sequence identity," as used herein, is determined by comparing two optimally aligned sequences over a comparison window, where the fragment of the polynucleotide or amino acid sequence in the comparison window may comprise additions or deletions (e.g., gaps or overhangs) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences.

The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Optimal alignment of sequences for comparison may be conducted by the local homology algorithm of Smith and Waterman *Add. APL. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman and Wunsch *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson and Lipman *Proc. Natl. Acad. Sci. (USA)* 85: 2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, BLAST, PASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group (GCG), 575 Science Dr., Madison, WI), or by inspection. Given that two sequences have been identified for comparison, GAP and BESTFIT are preferably employed to determine their optimal alignment. Typically, the default values of 5.00 for gap weight and 0.30 for gap weight length are used. The term "substantial sequence identity" between polynucleotide or polypeptide sequences refers to polynucleotide or polypeptide comprising a sequence that has at least 80% sequence identity, preferably at least 85%, more preferably at least 90% and most preferably at least 95%, even more preferably, at least 96%, 97%, 98% or 99% sequence identity compared to a reference sequence using the programs.

Plant Promoter A "plant promoter" is a promoter capable of initiating transcription in plant cells and can drive or facilitate transcription of a fragment of the SDF of the instant invention or a coding sequence of the SDF of the instant invention. Such promoters need not be of plant origin. For example, promoters derived from plant viruses, such as the CaMV35S promoter or from *Agrobacterium tumefaciens* such as the T-DNA promoters, can be plant promoters. A typical example of a plant promoter of plant origin is the maize ubiquitin-1 (ubi-1)promoter known to those of skill.

Promoter: The term "promoter," as used herein, refers to a region of sequence determinants located upstream from the start of transcription of a gene and which are involved in recognition and binding of RNA polymerase and other proteins to initiate and modulate transcription. A basal promoter is the minimal sequence necessary for assembly of a transcription complex required for transcription initiation. Basal promoters frequently include a

"TATA box" element usually located between 15 and 35 nucleotides upstream from the site of initiation of transcription. Basal promoters also sometimes include a "CCAAT box" element (typically a sequence CCAAT) and/or a GGGCG sequence, usually located between 40 and 200 nucleotides, preferably 60 to 120 nucleotides, upstream from the start site of transcription.

Public sequence: The term "public sequence ," as used in the context of the instant application, refers to any sequence that has been deposited in a publicly accessible database. This term encompasses both amino acid and nucleotide sequences. Such sequences are publicly accessible, for example, on the BLAST databases on the NCBI FTP web site (accessible at ncbi.nlm.gov/blast). The database at the NCBI GTP site utilizes "gi" numbers assigned by NCBI as a unique identifier for each sequence in the databases, thereby providing a non-redundant database for sequence from various databases, including GenBank, EMBL, DBBJ, (DNA Database of Japan) and PDB (Brookhaven Protein Data Bank).

Regulatory Sequence The term "regulatory sequence," as used in the current invention, refers to any nucleotide sequence that influences transcription or translation initiation and rate, and stability and/or mobility of the transcript or polypeptide product. Regulatory sequences include, but are not limited to, promoters, promoter control elements, protein binding sequences, 5' and 3' UTRs, transcriptional start site, termination sequence, polyadenylation sequence, introns, certain sequences within a coding sequence, etc.

Related Sequences: "Related sequences" refer to either a polypeptide or a nucleotide sequence that exhibits some degree of sequence similarity with a sequence described in Table 1.

Scaffold Attachment Region (SAR) As used herein, "scaffold attachment region" is a DNA sequence that anchors chromatin to the nuclear matrix or scaffold to generate loop domains that can have either a transcriptionally active or inactive structure (Spiker and Thompson (1996) Plant Physiol. 110: 15-21).

Sequence-determined DNA fragments (SDFs) "Sequence-determined DNA fragments" as used in the current invention are isolated sequences of genes, fragments of genes,

intergenic regions or contiguous DNA from plant genomic DNA or cDNA or RNA the sequence of which has been determined.

Signal Peptide A "signal peptide" as used in the current invention is an amino acid sequence that targets the protein for secretion, for transport to an intracellular compartment or organelle or for incorporation into a membrane. Signal peptides are indicated in the tables and a more detailed description located below.

Specific Promoter In the context of the current invention, "specific promoters" refers to a subset of inducible promoters that have a high preference for being induced in a specific tissue or cell and/or at a specific time during development of an organism. By "high preference" is meant at least 3-fold, preferably 5-fold, more preferably at least 10-fold still more preferably at least 20-fold, 50-fold or 100-fold increase in transcription in the desired tissue over the transcription in any other tissue. Typical examples of temporal and/or tissue specific promoters of plant origin that can be used with the polynucleotides of the present invention, are: PTA29, a promoter which is capable of driving gene transcription specifically in tapetum and only during anther development (Koltonow et al., *Plant Cell* 2:1201 (1990); RCc2 and RCc3, promoters that direct root-specific gene transcription in rice (Xu et al., *Plant Mol. Biol.* 27:237 (1995); TobRB27, a root-specific promoter from tobacco (Yamamoto et al., *Plant Cell* 3:371 (1991)). Examples of tissue-specific promoters under developmental control include promoters that initiate transcription only in certain tissues or organs, such as root, ovule, fruit, seeds, or flowers. Other suitable promoters include those from genes encoding storage proteins or the lipid body membrane protein, oleosin. A few root-specific promoters are noted above.

Stringency "Stringency" as used herein is a function of probe length, probe composition (G + C content), and salt concentration, organic solvent concentration, and temperature of hybridization or wash conditions. Stringency is typically compared by the parameter $T_m$, which is the temperature at which 50% of the complementary molecules in the hybridization are hybridized, in terms of a temperature differential from $T_m$. High stringency conditions are those providing a condition of $T_m$ - 5°C to $T_m$ - 10°C. Medium or moderate stringency conditions are those providing $T_m$ - 20°C to $T_m$ - 29°C. Low stringency conditions are those providing a condition of $T_m$ - 40°C to $T_m$ - 48°C. The relationship of hybridization conditions to $T_m$ (in °C) is expressed in the mathematical equation

$$T_m = 81.5 - 16.6(\log_{10}[Na^+]) + 0.41(\%G+C) - (600/N) \quad (1)$$

where N is the length of the probe. This equation works well for probes 14 to 70 nucleotides in length that are identical to the target sequence. The equation below for $T_m$ of DNA-DNA hybrids is useful for probes in the range of 50 to greater than 500 nucleotides, and for conditions that include an organic solvent (formamide).

$$T_m = 81.5+16.6 \log \{[Na^+]/(1+0.7[Na^+])\}+ 0.41(\%G+C)-500/L\ 0.63(\%formamide) \quad (2)$$

where L is the length of the probe in the hybrid. (P. Tijessen, "Hybridization with Nucleic Acid Probes" in Laboratory Techniques in Biochemistry and Molecular Biology, P.C. vand der Vliet, ed., c. 1993 by Elsevier, Amsterdam.) The $T_m$ of equation (2) is affected by the nature of the hybrid; for DNA-RNA hybrids $T_m$ is 10-15°C higher than calculated, for RNA-RNA hybrids $T_m$ is 20-25°C higher. Because the $T_m$ decreases about 1 °C for each 1% decrease in homology when a long probe is used (Bonner et al., *J. Mol. Biol.* 81:123 (1973)), stringency conditions can be adjusted to favor detection of identical genes or related family members.

Equation (2) is derived assuming equilibrium and therefore, hybridizations according to the present invention are most preferably performed under conditions of probe excess and for sufficient time to achieve equilibrium. The time required to reach equilibrium can be shortened by inclusion of a hybridization accelerator such as dextran sulfate or another high volume polymer in the hybridization buffer.

Stringency can be controlled during the hybridization reaction or after hybridization has occurred by altering the salt and temperature conditions of the wash solutions used. The formulas shown above are equally valid when used to compute the stringency of a wash solution. Preferred wash solution stringencies lie within the ranges stated above; high stringency is 5-8°C below $T_m$, medium or moderate stringency is 26-29°C below $T_m$ and low stringency is 45-48°C below $T_m$.

Substantially free of A composition containing A is "substantially free of " B when at least 85% by weight of the total A+B in the composition is A. Preferably, A comprises at least about 90% by weight of the total of A+B in the composition, more preferably at least about 95% or even 99% by weight. For example, a plant gene or DNA sequence can be considered substantially free of other plant genes or DNA sequences.

Translational start site In the context of the current invention, a "translational start site" is usually an ATG in the cDNA transcript, more usually the first ATG. A single cDNA, however, may have multiple translational start sites.

Transcription start site "Transcription start site" is used in the current invention to describe the point at which transcription is initiated. This point is typically located about 25 nucleotides downstream from a TFIID binding site, such as a TATA box. Transcription can initiate at one or more sites within the gene, and a single gene may have multiple transcriptional start sites, some of which may be specific for transcription in a particular cell-type or tissue.

Untranslated region (UTR) A "UTR" is any contiguous series of nucleotide bases that is transcribed, but is not translated. These untranslated regions may be associated with particular functions such as increasing mRNA message stability. Examples of UTRs include, but are not limited to polyadenylation signals, terminations sequences, sequences located between the transcriptional start site and the first exon (5' UTR) and sequences located between the last exon and the end of the mRNA (3' UTR).

Variant: The term "variant" is used herein to denote a polypeptide or protein or polynucleotide molecule that differs from others of its kind in some way. For example, polypeptide and protein variants can consist of changes in amino acid sequence and/or charge and/or post-translational modifications (such as glycosylation, etc).

## DETAILED DESCRIPTION OF THE INVENTION

### I. Polynucleotides

Exemplified SDFs of the invention represent fragments of the genome of corn, wheat, rice, soybean or *Arabidopsis* and/or represent mRNA expressed from that genome. The isolated nucleic acid of the invention also encompasses corresponding fragments of the genome and/or cDNA complement of other organisms as described in detail below.

Polynucleotides of the invention can be isolated from polynucleotide libraries using primers comprising sequence similar to those described by Table 1. See, for example, the methods described in Sambrook et al., supra.

Alternatively, the polynucleotides of the invention can be produced by chemical synthesis. Such synthesis methods are described below.

It is contemplated that the nucleotide sequences presented herein may contain some small percentage of errors. These errors may arise in the normal course of determination of nucleotide sequences. Sequence errors can be corrected by obtaining seeds deposited under the accession numbers cited herein, propagating them, isolating genomic DNA or appropriate mRNA from the resulting plants or seeds thereof, amplifying the relevant fragment of the genomic DNA or mRNA using primers having a sequence that flanks the erroneous sequence, and sequencing the amplification product.

I.A. Probes, Primers and Substrates

SDFs of the invention can be applied to substrates for use in array applications such as, but not limited to, assays of global gene expression, for example under varying conditions of development, growth conditions. The arrays can also be used in diagnostic or forensic methods (WO95/35505, US 5,445,943 and US 5,410,270).

Probes and primers of the instant invention will hybridize to a polynucleotide comprising a sequence in Table 1. Though many different nucleotide sequences can encode an amino acid sequence, the sequences of Table 1 are generally preferred for encoding polypeptides of the invention. However, the sequence of the probes and/or primers of the instant invention need not be identical to those in Table 1 or the complements thereof. For example, some variation in probe or primer sequence and/or length can allow additional family members to be detected, as well as orthologous genes and more taxonomically distant related sequences. Similarly, probes and/or primers of the invention can include additional nucleotides that serve as a label for detecting the formed duplex or for subsequent cloning purposes.

Probe length will vary depending on the application. For use as primers, probes are 12-40 nucleotides, preferably 18-30 nucleotides long. For use in mapping, probes are preferably 50 to 500 nucleotides, preferably 100-250 nucleotides long. For Southern hybridizations, probes as long as several kilobases can be used as explained below.

The probes and/or primers can be produced by synthetic procedures such as the triester method of Matteucci et al. *J. Am. Chem. Soc.* 103:3185( 1981); or according to Urdea

et al. *Proc. Natl. Acad.* 80:7461 (1981) or using commercially available automated oligonucleotide synthesizers.

I.B. Methods of Detection and Isolation

The polynucleotides of the invention can be utilized in a number of methods known to those skilled in the art as probes and/or primers to isolate and detect polynucleotides, including, without limitation: Southerns, Northerns, Branched DNA hybridization assays, polymerase chain reaction, and microarray assays, and variations thereof. Specific methods given by way of examples, and discussed below include:

Hybridization

Methods of Mapping

Southern Blotting

Isolating cDNA from Related Organisms

Isolating and/or Identifying Orthologous Genes.

Also, the nucleic acid molecules of the invention can used in other methods, such as high density oligonucleotide hybridizing assays, described, for example, in U.S. Pat. Nos. 6,004,753; 5,945,306; 5,945,287; 5,945,308; 5,919,686; 5,919,661; 5,919,627; 5,874,248; 5,871,973; 5,871,971; and 5,871,930; and PCT Pub. Nos. WO 9946380; WO 9933981; WO 9933870; WO 9931252; WO 9915658; WO 9906572; WO 9858052; WO 9958672; and WO 9810858.

B.1. Hybridization

The isolated SDFs of Table 1 of the present invention can be used as probes and/or primers for detection and/or isolation of related polynucleotide sequences through hybridization. Hybridization of one nucleic acid to another constitutes a physical property that defines the subject SDF of the invention and the identified related sequences. Also, such hybridization imposes structural limitations on the pair. A good general discussion of the factors for determining hybridization conditions is provided by Sambrook et al. ("Molecular Cloning, a Laboratory Manual, 2nd ed., c. 1989 by Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY; *see esp.*, chapters 11 and 12). Additional considerations and details of the physical chemistry of hybridization are provided by G.H. Keller and M.M. Manak "DNA Probes", 2nd Ed. pp. 1-25, c. 1993 by Stockton Press, New York, NY.

Depending on the stringency of the conditions under which these probes and/or primers are used, polynucleotides exhibiting a wide range of similarity to those in Table 1 can be detected or isolated. When the practitioner wishes to examine the result of membrane hybridizations under a variety of stringencies, an efficient way to do so is to perform the hybridization under a low stringency condition, then to wash the hybridization membrane under increasingly stringent conditions.

When using SDFs to identify orthologous genes in other species, the practitioner will preferably adjust the amount of target DNA of each species so that, as nearly as is practical, the same number of genome equivalents are present for each species examined. This prevents faint signals from species having large genomes, and thus small numbers of genome equivalents per mass of DNA, from erroneously being interpreted as absence of the corresponding gene in the genome.

The probes and/or primers of the instant invention can also be used to detect or isolate nucleotides that are "identical" to the probes or primers. Two nucleic acid sequences or polypeptides are said to be "identical" if the sequence of nucleotides or amino acid residues, respectively, in the two sequences is the same when aligned for maximum correspondence as described below.

Isolated polynucleotides within the scope of the invention also include allelic variants of the specific sequences presented in Table 1. The probes and/or primers of the invention can also be used to detect and/or isolate polynucleotides exhibiting at least 80% sequence identity with the sequences of Table 1 or fragments thereof.

With respect to nucleotide sequences, degeneracy of the genetic code provides the possibility to substitute at least one base of the base sequence of a gene with a different base without causing the amino acid sequence of the polypeptide produced from the gene to be changed. Hence, the DNA of the present invention may also have any base sequence that has been changed from a sequence in Table 1 by substitution in accordance with degeneracy of genetic code. References describing codon usage include: Carels *et al.*, *J. Mol. Evol.* 46: 45 (1998) and Fennoy *et al.*, *Nucl. Acids Res.* 21(23): 5294 (1993).

B.2. Mapping

The isolated SDF DNA of the invention can be used to create various types of genetic and physical maps of the genome of corn, Arabidopsis, soybean, rice, wheat, or other plants.

Some SDFs may be absolutely associated with particular phenotypic traits, allowing construction of gross genetic maps. While not all SDFs will immediately be associated with a phenotype, all SDFs can be used as probes for identifying polymorphisms associated with phenotypes of interest. Briefly, one method of mapping involves total DNA isolation from individuals. It is subsequently cleaved with one or more restriction enzymes, separated according to mass, transferred to a solid support, hybridized with SDF DNA and the pattern of fragments compared. Polymorphisms associated with a particular SDF are visualized as differences in the size of fragments produced between individual DNA samples after digestion with a particular restriction enzyme and hybridization with the SDF. After identification of polymorphic SDF sequences, linkage studies can be conducted. By using the individuals showing polymorphisms as parents in crossing programs, F2 progeny recombinants or recombinant inbreds, for example, are then analyzed. The order of DNA polymorphisms along the chromosomes can be determined based on the frequency with which they are inherited together versus independently. The closer two polymorphisms are together in a chromosome the higher the probability that they are inherited together. Integration of the relative positions of all the polymorphisms and associated marker SDFs can produce a genetic map of the species, where the distances between markers reflect the recombination frequencies in that chromosome segment.

The use of recombinant inbred lines for such genetic mapping is described for *Arabidopsis* by Alonso-Blanco et al. (*Methods in Molecular Biology*, vol.82, "*Arabidopsis Protocols*", pp. 137-146, J.M. Martinez-Zapater and J. Salinas, eds., c. 1998 by Humana Press, Totowa, NJ) and for corn by Burr ("Mapping Genes with Recombinant Inbreds", pp. 249-254. *In* Freeling, M. and V. Walbot (Ed.), *The Maize Handbook*, c. 1994 by Springer-Verlag New York, Inc.: New York, NY, USA; Berlin Germany; Burr et al. *Genetics* (1998) 118: 519; Gardiner, J. et al., (1993) *Genetics* 134: 917). This procedure, however, is not limited to plants and can be used for other organisms (such as yeast) or for individual cells.

The SDFs of the present invention can also be used for simple sequence repeat (SSR) mapping. Rice SSR mapping is described by Morgante et al. (*The Plant Journal* (1993) 3: 165), Panaud et al. (*Genome* (1995) 38: 1170); Senior et al. (*Crop Science* (1996) 36: 1676), Taramino et al. (*Genome* (1996) 39: 277) and Ahn et al. (*Molecular and General Genetics* (1993) 241: 483-90). SSR mapping can be achieved using various methods. In one instance, polymorphisms are identified when sequence specific probes contained within an SDF flanking an SSR are made and used in polymerase chain reaction (PCR) assays with template DNA from two or more individuals of interest. Here, a change in the number of tandem

repeats between the SSR-flanking sequences produces differently sized fragments (U.S. Patent 5,766,847). Alternatively, polymorphisms can be identified by using the PCR fragment produced from the SSR-flanking sequence specific primer reaction as a probe against Southern blots representing different individuals (U.H. Refseth et al., (1997) *Electrophoresis* 18: 1519).

Genetic and physical maps of crop species have many uses. For example, these maps can be used to devise positional cloning strategies for isolating novel genes from the mapped crop species. In addition, because the genomes of closely related species are largely syntenic (that is, they display the same ordering of genes within the genome), these maps can be used to isolate novel alleles from relatives of crop species by positional cloning strategies.

The various types of maps discussed above can be used with the SDFs of the invention to identify Quantitative Trait Loci (QTLs). Many important crop traits, such as the solids content of tomatoes, are quantitative traits and result from the combined interactions of several genes. These genes reside at different loci in the genome, oftentimes on different chromosomes, and generally exhibit multiple alleles at each locus. The SDFs of the invention can be used to identify QTLs and isolate specific alleles as described by de Vicente and Tanksley (*Genetics* 134:585 (1993)). In addition to isolating QTL alleles in present crop species, the SDFs of the invention can also be used to isolate alleles from the corresponding QTL of wild relatives. Transgenic plants having various combinations of QTL alleles can then be created and the effects of the combinations measured. Once a desired allele combination has been identified, crop improvement can be accomplished either through biotechnological means or by directed conventional breeding programs (for review *see* Tanksley and McCouch, *Science* 277:1063 (1997)).

In another embodiment, the SDFs can be used to help create physical maps of the genome of corn, *Arabidopsis* and related species. Where SDFs have been ordered on a genetic map, as described above, they can be used as probes to discover which clones in large libraries of plant DNA fragments in YACs, BACs, etc. contain the same SDF or similar sequences, thereby facilitating the assignment of the large DNA fragments to chromosomal positions. Subsequently, the large BACs, YACs, etc. can be ordered unambiguously by more detailed studies of their sequence composition (e.g. Marra et al. (1997) Genomic Research 7:1072-1084) and by using their end or other sequences to find the identical sequences in other cloned DNA fragments. The overlapping of DNA sequences in this way allows large contigs of plant sequences to be built that, when sufficiently extended, provide a complete

physical map of a chromosome. Sometimes the SDFs themselves will provide the means of joining cloned sequences into a contig.

The patent publication WO95/35505 and U.S. Patents 5,445,943 and 5,410,270 describe scanning multiple alleles of a plurality of loci using hybridization to arrays of oligonucleotides. These techniques are useful for each of the types of mapping discussed above.

Following the procedures described above and using a plurality of the SDFs of the present invention, any individual can be genotyped. These individual genotypes can be used for the identification of particular cultivars, varieties, lines, ecotypes and genetically modified plants or can serve as tools for subsequent genetic studies involving multiple phenotypic traits.

B.3 Southern Blot Hybridization

The sequences from Table 1 can be used as probes for various hybridization techniques. These techniques are useful for detecting target polynucleotides in a sample or for determining whether transgenic plants, seeds or host cells harbor a gene or sequence of interest and thus might be expected to exhibit a particular trait or phenotype.

In addition, the SDFs from the invention can be used to isolate additional members of gene families from the same or different species and/or orthologous genes from the same or different species. This is accomplished by hybridizing an SDF to, for example, a Southern blot containing the appropriate genomic DNA or cDNA. Given the resulting hybridization data, one of ordinary skill in the art could distinguish and isolate the correct DNA fragments by size, restriction sites, sequence and stated hybridization conditions from a gel or from a library.

Identification and isolation of orthologous genes from closely related species and alleles within a species is particularly desirable because of their potential for crop improvement. Many important crop traits, such as the solid content of tomatoes, result from the combined interactions of the products of several genes residing at different loci in the genome. Generally, alleles at each of these loci can make quantitative differences to the trait. By identifying and isolating numerous alleles for each locus from within or different species, transgenic plants with various combinations of alleles can be created and the effects of the combinations measured. Once a more favorable allele combination has been identified, crop improvement can be accomplished either through biotechnological means or by directed conventional breeding programs (Tanksley et al. *Science* 277:1063(1997)).

The results from hybridizations of the SDFs of the invention to, for example, Southern blots containing DNA from another species can also be used to generate restriction fragment maps for the corresponding genomic regions. These maps provide additional information about the relative positions of restriction sites within fragments, further distinguishing mapped DNA from the remainder of the genome.

Physical maps can be made by digesting genomic DNA with different combinations of restriction enzymes.

Probes for Southern blotting to distinguish individual restriction fragments can range in size from 15 to 20 nucleotides to several thousand nucleotides. More preferably, the probe is 100 to 1,000 nucleotides long for identifying members of a gene family when it is found that repetitive sequences would complicate the hybridization. For identifying an entire corresponding gene in another species, the probe is more preferably the length of the gene, typically 2,000 to 10,000 nucleotides, but probes 50-1,000 nucleotides long might be used. Some genes, however, might require probes up to 1,500 nucleotides long or overlapping probes constituting the full-length sequence to span their lengths.

Also, while it is preferred that the probe be homogeneous with respect to its sequence, it is not necessary. For example, as described below, a probe representing members of a gene family having diverse sequences can be generated using PCR to amplify genomic DNA or RNA templates using primers derived from SDFs that include sequences that define the gene family.

For identifying corresponding genes in another species, the next most preferable probe is a cDNA spanning the entire coding sequence, which allows all of the mRNA-coding fragment of the gene to be identified. Probes for Southern blotting can easily be generated from SDFs by making primers having the sequence at the ends of the SDF and using corn or *Arabidopsis* genomic DNA as a template. In instances where the SDF includes sequence conserved among species, primers including the conserved sequence can be used for PCR with genomic DNA from a species of interest to obtain a probe.

Similarly, if the SDF includes a domain of interest, that fragment of the SDF can be used to make primers and, with appropriate template DNA, used to make a probe to identify genes containing the domain. Alternatively, the PCR products can be resolved, for example by gel electrophoresis, and cloned and/or sequenced. Using Southern hybridization, the variants of the domain among members of a gene family, both within and across species, can be examined.

B.4.1 Isolating DNA from Related Organisms

The SDFs of the invention can be used to isolate the corresponding DNA from other organisms. Either cDNA or genomic DNA can be isolated. For isolating genomic DNA, a lambda, cosmid, BAC or YAC, or other large insert genomic library from the plant of interest can be constructed using standard molecular biology techniques as described in detail by Sambrook et al. 1989 (Molecular Cloning: A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory Press, New York) and by Ausubel et al. 1992 (Current Protocols in Molecular Biology, Greene Publishing, New York).

To screen a phage library, for example, recombinant lambda clones are plated out on appropriate bacterial medium using an appropriate *E. coli* host strain. The resulting plaques are lifted from the plates using nylon or nitrocellulose filters. The plaque lifts are processed through denaturation, neutralization, and washing treatments following the standard protocols outlined by Ausubel et al. (1992). The plaque lifts are hybridized to either radioactively labeled or non-radioactively labeled SDF DNA at room temperature for about 16 hours, usually in the presence of 50% formamide and 5X SSC (sodium chloride and sodium citrate) buffer and blocking reagents. The plaque lifts are then washed at 42°C with 1% Sodium Dodecyl Sulfate (SDS) and at a particular concentration of SSC. The SSC concentration used is dependent upon the stringency at which hybridization occurred in the initial Southern blot analysis performed. For example, if a fragment hybridized under medium stringency (e.g., Tm - 20°C), then this condition is maintained or preferably adjusted to a less stringent condition (e.g., Tm-30°C) to wash the plaque lifts. Positive clones show detectable hybridization e.g., by exposure to X-ray films or chromogen formation. The positive clones are then subsequently isolated for purification using the same general protocol outlined above. Once the clone is purified, restriction analysis can be conducted to narrow the region corresponding to the gene of interest. The restriction analysis and succeeding subcloning steps can be done using procedures described by, for example Sambrook et al. (1989) cited above.

The procedures outlined for the lambda library are essentially similar to those used for YAC library screening, except that the YAC clones are harbored in bacterial colonies. The YAC clones are plated out at reasonable density on nitrocellulose or nylon filters supported by appropriate bacterial medium in petri plates. Following the growth of the bacterial clones, the filters are processed through the denaturation, neutralization, and washing steps following

the procedures of Ausubel et al. 1992. The same hybridization procedures for lambda library screening are followed.

To isolate cDNA, similar procedures using appropriately modified vectors are employed. For instance, the library can be constructed in a lambda vector appropriate for cloning cDNA such as λgt11. Alternatively, the cDNA library can be made in a plasmid vector. cDNA for cloning can be prepared by any of the methods known in the art, but is preferably prepared as described above. Preferably, a cDNA library will include a high proportion of full-length clones.

B. 5. Isolating and/or Identifying Orthologous Genes

Probes and primers of the invention can be used to identify and/or isolate polynucleotides related to those in Table 1. Related polynucleotides are those that are native to other plant organisms and exhibit either similar sequence or encode polypeptides with similar biological activity. One specific example is an orthologous gene. Orthologous genes have the same functional activity. As such, orthologous genes may be distinguished from homologous genes. The percentage of identity is a function of evolutionary separation and, in closely related species, the percentage of identity can be 98 to 100%. The amino acid sequence of a protein encoded by an orthologous gene can be less than 75% identical, but tends to be at least75% or at least 80% identical, more preferably at least 90%, most preferably at least 95% identical to the amino acid sequence of the reference protein.

To find orthologous genes, the probes are hybridized to nucleic acids from a species of interest under low stringency conditions, preferably one where sequences containing as much as 40-45% mismatches will be able to hybridize. This condition is established by $T_m$ - 40°C to Tm - 48°C (*see* below). Blots are then washed under conditions of increasing stringency. It is preferable that the wash stringency be such that sequences that are 85 to 100% identical will hybridize. More preferably, sequences 90 to 100% identical will hybridize and most preferably only sequences greater than 95% identical will hybridize. One of ordinary skill in the art will recognize that, due to degeneracy in the genetic code, amino acid sequences that are identical can be encoded by DNA sequences as little as 67% identical or less. Thus, it is preferable, for example, to make an overlapping series of shorter probes, on the order of 24 to 45 nucleotides, and individually hybridize them to the same arrayed library to avoid the problem of degeneracy introducing large numbers of mismatches.

As evolutionary divergence increases, genome sequences also tend to diverge. Thus, one of skill will recognize that searches for orthologous genes between more divergent species will require the use of lower stringency conditions compared to searches between closely related species. Also, degeneracy of the genetic code is more of a problem for searches in the genome of a species more distant evolutionarily from the species that is the source of the SDF probe sequences.

Therefore the method described in Bouckaert et al., U.S. Ser. No. 60/121,700 Atty. Dkt. No. 2750-117P, Client Dkt. No. 00010.001, filed February 25, 1999, hereby incorporated in its entirety by reference, can be applied to the SDFs of the present invention to isolate related genes from plant species which do not hybridize to the corn *Arabidopsis*, soybean, rice, wheat, and other plant sequences of Table 1.

Identification of the relationship of nucleotide or amino acid sequences among plant species can be done by comparing the nucleotide or amino acid sequences of SDFs of the present application with nucleotide or amino acid sequences of other SDFs such as those present in applications listed in the table below:

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0301P | 80002.001 | 9/4/98 | 60/099,672 |
| United States | 2750-0300P | 80001.001 | 9/4/98 | 60/099,671 |
| United States | 2750-0302P | 80003.001 | 9/11/98 | 60/099,933 |
| United States | 2750-0304P | 80004.001 | 9/17/98 | 60/100,864 |
| United States | 2750-0305P | 80005.001 | 9/18/98 | 60/101,042 |
| United States | 2750-0306P | 80006.001 | 9/21/98 | 60/101,255 |
| United States | 2750-0307P | 80007.001 | 9/24/98 | 60/101,682 |
| United States | 2750-0308P | 80008.001 | 9/30/98 | 60/102,533 |
| United States | 2750-0309P | 80009.001 | 9/30/98 | 60/102,460 |
| United States | 2750-0310P | 80010.001 | 10/5/98 | 60/103,116 |
| United States | 2750-0311P | 80011.001 | 10/5/98 | 60/103,141 |
| United States | 2750-0312P | 80012.001 | 10/6/98 | 60/103,215 |
| United States | 2750-0313P | 80013.001 | 10/8/98 | 60/103,554 |
| United States | 2750-0314P | 80014.001 | 10/9/98 | 60/103,574 |
| United States | 2750-0315P | 80015.001 | 10/13/98 | 60/103,907 |
| United States | 2750-0316P | 80016.001 | 10/14/98 | 60/104,268 |
| United States | 2750-0317P | 80017.001 | 10/16/98 | 60/104,680 |
| United States | 2750-0318P | 80018.001 | 10/19/98 | 60/104,828 |
| United States | 2750-0319P | 80019.001 | 10/20/98 | 60/105,008 |
| United States | 2750-0320P | 80020.001 | 10/21/98 | 60/105,142 |
| United States | 2750-0321P | 80021.001 | 10/22/98 | 60/105,533 |
| United States | 2750-0322P | 80022.001 | 10/26/98 | 60/105,571 |
| United States | 2750-0323P | 80023.001 | 10/27/98 | 60/105,815 |
| United States | 2750-0324P | 80024.001 | 10/29/98 | 60/106,105 |
| United States | 2750-0325P | 80025.001 | 10/30/98 | 60/106,218 |
| United States | 2750-0326P | 80026.001 | 11/2/98 | 60/106,685 |
| United States | 2750-0327P | 80027.001 | 11/6/98 | 60/107,282 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0328P | 80028.001 | 11/9/98 | 60/107,720 |
| United States | 2750-0329P | 80029.001 | 11/9/98 | 60/107,719 |
| United States | 2750-0330P | 80030.001 | 11/10/98 | 60/107,836 |
| United States | 2750-0331P | 80031.001 | 11/12/98 | 60/108,190 |
| United States | 2750-0332P | 80032.001 | 11/16/98 | 60/108,526 |
| United States | 2750-0333P | 80033.001 | 11/17/98 | 60/108,901 |
| United States | 2750-0335P | 80035.001 | 11/19/98 | 60/109,127 |
| United States | 2750-0334P | 80034.001 | 11/19/98 | 60/109,124 |
| United States | 2750-0336P | 80036.001 | 11/20/98 | 60/109,267 |
| United States | 2750-0337P | 80037.001 | 11/23/98 | 60/109,594 |
| United States | 2750-0339P | 80039.001 | 11/25/98 | 60/110,050 |
| United States | 2750-0338P | 80038.001 | 11/25/98 | 60/110,053 |
| United States | 2750-0340P | 80040.001 | 11/27/98 | 60/110,158 |
| United States | 2750-0341P | 80041.001 | 11/30/98 | 60/110,263 |
| United States | 2750-0342P | 80042.001 | 12/1/98 | 60/110,495 |
| United States | 2750-0343P | 80043.001 | 12/2/98 | 60/110,626 |
| United States | 2750-0344P | 80044.001 | 12/3/98 | 60/110,701 |
| United States | 2750-0345P | 80045.001 | 12/7/98 | 60/111,339 |
| United States | 2750-0346P | 80046.001 | 12/9/98 | 60/111,589 |
| United States | 2750-0347P | 80047.001 | 12/10/98 | 60/111,782 |
| United States | 2750-0348P | 80048.001 | 12/11/98 | 60/111,812 |
| United States | 2750-0349P | 80049.001 | 12/14/98 | 60/112,096 |
| United States | 2750-0350P | 80050.001 | 12/15/98 | 60/112,224 |
| United States | 2750-0351P | 80051.001 | 12/16/98 | 60/112,624 |
| United States | 2750-0352P | 80052.001 | 12/17/98 | 60/112,862 |
| United States | 2750-0353P | 80053.001 | 12/18/98 | 60/112,912 |
| United States | 2750-0354P | 80054.001 | 12/21/98 | 60/113,248 |
| United States | 2750-0355P | 80055.001 | 12/22/98 | 60/113,522 |
| United States | 2750-0356P | 80056.001 | 12/23/98 | 60/113,826 |
| United States | 2750-0357P | 80057.001 | 12/28/98 | 60/113,998 |
| United States | 2750-0358P | 80058.001 | 12/29/98 | 60/114,384 |
| United States | 2750-0359P | 80059.001 | 12/30/98 | 60/114,455 |
| United States | 2750-0360P | 80060.001 | 1/4/99 | 60/114,740 |
| United States | 2750-0361P | 80061.001 | 1/6/99 | 60/114,866 |
| United States | 2750-0364P | 80064.001 | 1/7/99 | 60/115,151 |
| United States | 2750-0363P | 80063.001 | 1/7/99 | 60/115,152 |
| United States | 2750-0367P | 80067.001 | 1/7/99 | 60/115,154 |
| United States | 2750-0366P | 80066.001 | 1/7/99 | 60/115,156 |
| United States | 2750-0365P | 80065.001 | 1/7/99 | 60/115,155 |
| United States | 2750-0362P | 80062.001 | 1/7/99 | 60/115,153 |
| United States | 2750-0370P | 80070.001 | 1/8/99 | 60/115,293 |
| United States | 2750-0369P | 80069.001 | 1/8/99 | 60/115,365 |
| United States | 2750-0368P | 80068.001 | 1/8/99 | 60/115,364 |
| United States | 2750-0371P | 80071.001 | 1/11/99 | 60/115,339 |
| United States | 2750-0372P | 80072.001 | 1/12/99 | 60/115,518 |
| United States | 2750-0373P | 80073.001 | 1/13/99 | 60/115,847 |
| United States | 2750-0374P | 80074.001 | 1/14/99 | 60/115,905 |
| United States | 2750-0375P | 80075.001 | 1/15/99 | 60/116,383 |
| United States | 2750-0376P | 80076.001 | 1/15/99 | 60/116,384 |
| United States | 2750-0378P | 80078.001 | 1/19/99 | 60/116,340 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0377P | 80077.001 | 1/19/99 | 60/116,329 |
| United States | 2750-0380P | 80080.001 | 1/21/99 | 60/116,672 |
| United States | 2750-0379P | 80079.001 | 1/21/99 | 60/116,674 |
| United States | 2750-0382P | 80082.001 | 1/22/99 | 60/116,962 |
| United States | 2750-0381P | 80081.001 | 1/22/99 | 60/116,960 |
| United States | 2750-0383P | 80083.001 | 1/28/99 | 60/117,756 |
| United States | 2750-0384P | 80084.001 | 2/3/99 | 60/118,672 |
| United States | 2750-0385P | 80085.001 | 2/4/99 | 60/118,808 |
| United States | 2750-0386P | 80086.001 | 2/5/99 | 60/118,778 |
| United States | 2750-0387P | 80087.001 | 2/8/99 | 60/119,029 |
| United States | 2750-0388P | 80088.001 | 2/9/99 | 60/119,332 |
| United States | 2750-0389P | 80089.001 | 2/10/99 | 60/119,462 |
| United States | 2750-0391P | 80091.001 | 2/12/99 | 60/119,922 |
| United States | 2750-0393P | 80093.001 | 2/16/99 | 60/120,198 |
| United States | 2750-0392P | 80092.001 | 2/16/99 | 60/120,196 |
| United States | 2750-0394P | 80094.001 | 2/18/99 | 60/120,583 |
| United States | 2750-0395P | 80095.001 | 2/22/99 | 60/121,072 |
| United States | 2750-0396P | 80096.001 | 2/23/99 | 60/121,334 |
| United States | 2750-0397P | 80097.001 | 2/24/99 | 60/121,470 |
| United States | 2750-0398P | 80098.001 | 2/25/99 | 60/121,704 |
| United States | 2750-0390P | 80090.001 | 2/25/99 | 60/121,825 |
| United States | 2750-0399P | 80099.001 | 2/26/99 | 60/122,107 |
| United States | 2750-0400P | 80100.001 | 3/1/99 | 60/122,266 |
| United States | 2750-0401P | 80101.001 | 3/2/99 | 60/122,568 |
| United States | 2750-0402P | 80102.001 | 3/3/99 | 60/122,611 |
| United States | 2750-0403P | 80103.001 | 3/4/99 | 60/121,775 |
| United States | 2750-0405P | 80105.001 | 3/5/99 | 60/123,180 |
| United States | 2750-0404P | 80104.001 | 3/5/99 | 60/123,534 |
| United States | 2750-0407P | 80107.001 | 3/9/99 | 60/123,548 |
| United States | 2750-0406P | 80106.001 | 3/9/99 | 60/123,680 |
| United States | 2750-0408P | 80108.001 | 3/10/99 | 60/123,715 |
| United States | 2750-0409P | 80109.001 | 3/10/99 | 60/123,726 |
| United States | 2750-0410P | 80110.001 | 3/11/99 | 60/124,263 |
| United States | 2750-0411P | 80111.001 | 3/12/99 | 60/123,941 |
| United States | 2750-0412P | 80112.001 | 3/23/99 | 60/125,788 |
| United States | 2750-0413P | 80113.001 | 3/25/99 | 60/126,264 |
| United States | 2750-0414P | 80114.001 | 3/29/99 | 60/126,785 |
| United States | 2750-0415P | 80115.001 | 4/1/99 | 60/127,462 |
| United States | 2750-0416P | 91000.001 | 4/6/99 | 60/128,234 |
| United States | 2750-0417P | 91001.001 | 4/8/99 | 60/128,714 |
| United States | 2750-0418P | 80118.001 | 4/16/99 | 60/129,845 |
| United States | 2750-0420P | 80120.001 | 4/19/99 | 60/130,077 |
| United States | 2750-0421P | 80121.001 | 4/21/99 | 60/130,449 |
| United States | 2750-0303P | 80115.002 | 4/23/99 | 60/130,510 |
| United States | 2750-0422P | 80122.001 | 4/23/99 | 60/130,891 |
| United States | 2750-0423P | 80123.001 | 4/28/99 | 60/131,449 |
| United States | 2750-0424P | 80124.001 | 4/30/99 | 60/132,407 |
| United States | 2750-0425P | 80125.001 | 4/30/99 | 60/132,048 |
| United States | 2750-0426P | 80126.001 | 5/4/99 | 60/132,484 |
| United States | 2750-0427P | 80127.001 | 5/5/99 | 60/132,485 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0428P | 91002.001 | 5/6/99 | 60/132,487 |
| United States | 2750-0429P | 80129.001 | 5/6/99 | 60/132,486 |
| United States | 2750-0430P | 80130.001 | 5/7/99 | 60/132,863 |
| United States | 2750-0431P | 80131.001 | 5/11/99 | 60/134,256 |
| United States | 2750-0432P | 91006.001 | 5/14/99 | 60/134,370 |
| United States | 2750-0434P | 80116.001 | 5/14/99 | 60/134,219 |
| United States | 2750-0435P | 80117.001 | 5/14/99 | 60/134,218 |
| United States | 2750-0433P | 00025.001 | 5/14/99 | 60/134,221 |
| United States | 2750-0436P | 91007.001 | 5/18/99 | 60/134,768 |
| United States | 2750-0437P | 91008.001 | 5/19/99 | 60/134,941 |
| United States | 2750-0438P | 91009.001 | 5/20/99 | 60/135,124 |
| United States | 2750-0439P | 91010.001 | 5/21/99 | 60/135,353 |
| United States | 2750-0440P | 91011.001 | 5/24/99 | 60/135,629 |
| United States | 2750-0441P | 91012.001 | 5/25/99 | 60/136,021 |
| United States | 2750-0442P | 91013.001 | 5/27/99 | 60/136,392 |
| United States | 2750-0444P | 91014.001 | 5/28/99 | 60/136,782 |
| United States | 2750-0445P | 91015.001 | 6/1/99 | 60/137,222 |
| United States | 2750-0446P | 91016.001 | 6/3/99 | 60/137,528 |
| United States | 2750-0447P | 91017.001 | 6/4/99 | 60/137,502 |
| United States | 2750-0449P | 91018.001 | 6/7/99 | 60/137,724 |
| United States | 2750-0450P | 91019.001 | 6/8/99 | 60/138,094 |
| United States | 2750-0458P | 00033.002 | 6/10/99 | 60/138,847 |
| United States | 2750-0457P | 00033.001 | 6/10/99 | 60/138,540 |
| United States | 2750-0463P | 00034.001 | 6/14/99 | 60/139,119 |
| United States | 2750-0462P | 80132.012 | 6/16/99 | 60/139,452 |
| United States | 2750-0461P | 80132.011 | 6/16/99 | 60/139,453 |
| United States | 2750-0464P | 00037.001 | 6/17/99 | 60/139,492 |
| United States | 2750-0451P | 80132.003 | 6/18/99 | 60/139,459 |
| United States | 2750-0466P | 00039.001 | 6/18/99 | 60/139,750 |
| United States | 2750-0460P | 80132.010 | 6/18/99 | 60/139,455 |
| United States | 2750-0465P | 00038.001 | 6/18/99 | 60/139,763 |
| United States | 2750-0456P | 80132.008 | 6/18/99 | 60/139,456 |
| United States | 2750-0455P | 80132.007 | 6/18/99 | 60/139,460 |
| United States | 2750-0454P | 80132.006 | 6/18/99 | 60/139,457 |
| United States | 2750-0452P | 80132.004 | 6/18/99 | 60/139,461 |
| United States | 2750-0453P | 80132.005 | 6/18/99 | 60/139,462 |
| United States | 2750-0448P | 80132.002 | 6/18/99 | 60/139,454 |
| United States | 2750-0443P | 80132.001 | 6/18/99 | 60/139,458 |
| United States | 2750-0459P | 80132.009 | 6/18/99 | 60/139,463 |
| United States | 2750-0467P | 00042.001 | 6/21/99 | 60/139,817 |
| United States | 2750-0468P | 00043.001 | 6/22/99 | 60/139,899 |
| United States | 2750-0469P | 00044.001 | 6/23/99 | 60/140,354 |
| United States | 2750-0470P | 00042.002 | 6/23/99 | 60/140,353 |
| United States | 2750-0471P | 00045.001 | 6/24/99 | 60/140,695 |
| United States | 2750-0472P | 00046.001 | 6/28/99 | 60/140,823 |
| United States | 2750-0473P | 00048.001 | 6/29/99 | 60/140,991 |
| United States | 2750-0474P | 00049.001 | 6/30/99 | 60/141,287 |
| United States | 2750-0475P | 00050.001 | 7/1/99 | 60/141,842 |
| United States | 2750-0476P | 00051.001 | 7/1/99 | 60/142,154 |
| United States | 2750-0477P | 00052.001 | 7/2/99 | 60/142,055 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0478P | 00053.001 | 7/6/99 | 60/142,390 |
| United States | 2750-0479P | 00054.001 | 7/8/99 | 60/142,803 |
| United States | 2750-0480P | 00058.001 | 7/9/99 | 60/142,920 |
| United States | 2750-0481P | 00059.001 | 7/12/99 | 60/142,977 |
| United States | 2750-0482P | 00060.001 | 7/13/99 | 60/143,542 |
| United States | 2750-0489P | 00061.001 | 7/14/99 | 60/143,624 |
| United States | 2750-0490P | 00062.001 | 7/15/99 | 60/144,005 |
| United States | 2750-0485P | 80134.003 | 7/16/99 | 60/144,086 |
| United States | 2750-0486P | 80134.004 | 7/16/99 | 60/144,085 |
| United States | 2750-0496P | 80134.014 | 7/19/99 | 60/144,334 |
| United States | 2750-0497P | 00064.001 | 7/19/99 | 60/144,325 |
| United States | 2750-0495P | 80134.013 | 7/19/99 | 60/144,335 |
| United States | 2750-0494P | 80134.010 | 7/19/99 | 60/144,333 |
| United States | 2750-0492P | 80134.008 | 7/19/99 | 60/144,331 |
| United States | 2750-0488P | 80134.006 | 7/19/99 | 60/144,332 |
| United States | 2750-0502P | 80135.002 | 7/20/99 | 60/144,884 |
| United States | 2750-0500P | 00065.001 | 7/20/99 | 60/144,632 |
| United States | 2750-0499P | 80134.012 | 7/20/99 | 60/144,352 |
| United States | 2750-0503P | 00066.001 | 7/21/99 | 60/144,814 |
| United States | 2750-0483P | 80134.001 | 7/21/99 | 60/145,088 |
| United States | 2750-0484P | 80134.002 | 7/21/99 | 60/145,086 |
| United States | 2750-0487P | 80134.005 | 7/22/99 | 60/145,089 |
| United States | 2750-0491P | 80134.007 | 7/22/99 | 60/145,085 |
| United States | 2750-0493P | 80134.009 | 7/22/99 | 60/145,087 |
| United States | 2750-0504P | 00067.001 | 7/22/99 | 60/145,192 |
| United States | 2750-0498P | 80134.011 | 7/23/99 | 60/145,145 |
| United States | 2750-0505P | 00069.001 | 7/23/99 | 60/145,218 |
| United States | 2750-0501P | 80135.001 | 7/23/99 | 60/145,224 |
| United States | 2750-0506P | 00070.001 | 7/26/99 | 60/145,276 |
| United States | 2750-0507P | 80136.001 | 7/27/99 | 60/145,918 |
| United States | 2750-0508P | 80136.002 | 7/27/99 | 60/145,919 |
| United States | 2750-0509P | 00071.001 | 7/27/99 | 60/145,913 |
| United States | 2750-0510P | 00072.001 | 7/28/99 | 60/145,951 |
| United States | 2750-0511P | 80137.001 | 8/2/99 | 60/146,388 |
| United States | 2750-0513P | 00073.001 | 8/2/99 | 60/146,386 |
| United States | 2750-0512P | 80137.002 | 8/2/99 | 60/146,389 |
| United States | 2750-0514P | 00074.001 | 8/3/99 | 60/147,038 |
| United States | 2750-0515P | 00076.001 | 8/4/99 | 60/147,204 |
| United States | 2750-0517P | 80138.002 | 8/4/99 | 60/147,302 |
| United States | 2750-0518P | 00077.001 | 8/5/99 | 60/147,260 |
| United States | 2750-0519P | 80136.003 | 8/5/99 | 60/147,192 |
| United States | 2750-0516P | 80138.001 | 8/6/99 | 60/147,303 |
| United States | 2750-0520P | 00079.001 | 8/6/99 | 60/147,416 |
| United States | 2750-0523P | 80139.002 | 8/9/99 | 60/147,935 |
| United States | 2750-0521P | 00080.001 | 8/9/99 | 60/147,493 |
| United States | 2750-0522P | 80139.001 | 8/10/99 | 60/148,171 |
| United States | 2750-0524P | 00081.001 | 8/11/99 | 60/148,319 |
| United States | 2750-0530P | 00082.001 | 8/12/99 | 60/148,341 |
| United States | 2750-0525P | 80141.001 | 8/12/99 | 60/148,347 |
| United States | 2750-0526P | 80141.002 | 8/12/99 | 60/148,342 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0527P | 80141.003 | 8/12/99 | 60/148,340 |
| United States | 2750-0528P | 80141.004 | 8/12/99 | 60/148,337 |
| United States | 2750-0532P | 80142.002 | 8/13/99 | 60/148,684 |
| United States | 2750-0529P | 00083.001 | 8/13/99 | 60/148,565 |
| United States | 2750-0531P | 80142.001 | 8/16/99 | 60/149,368 |
| United States | 2750-0533P | 80001.002 | 8/17/99 | 60/149,927 |
| United States | 2750-0534P | 80001.003 | 8/17/99 | 60/149,928 |
| United States | 2750-0535P | 80001.004 | 8/17/99 | 60/149,926 |
| United States | 2750-0536P | 80001.005 | 8/17/99 | 60/149,925 |
| United States | 2750-0537P | 00084.001 | 8/17/99 | 60/149,175 |
| United States | 2750-0538P | 00085.001 | 8/18/99 | 60/149,426 |
| United States | 2750-0542P | 00087.001 | 8/20/99 | 60/149,723 |
| United States | 2750-0539P | 00086.001 | 8/20/99 | 60/149,722 |
| United States | 2750-0541P | 80143.002 | 8/20/99 | 60/149,929 |
| United States | 2750-0540P | 80143.001 | 8/23/99 | 60/149,930 |
| United States | 2750-0543P | 00088.001 | 8/23/99 | 60/149,902 |
| United States | 2750-0544P | 00089.001 | 8/25/99 | 60/150,566 |
| United States | 2750-0547P | 00090.001 | 8/26/99 | 60/150,884 |
| United States | 2750-0548P | 00091.001 | 8/27/99 | 60/151,080 |
| United States | 2750-0545P | 80144.001 | 8/27/99 | 60/151,065 |
| United States | 2750-0546P | 80144.002 | 8/27/99 | 60/151,066 |
| United States | 2750-0549P | 00092.001 | 8/30/99 | 60/151,303 |
| United States | 2750-0552P | 00093.001 | 8/31/99 | 60/151,438 |
| United States | 2750-0553P | 00094.001 | 9/1/99 | 60/151,930 |
| International | 2750-0551F(PC) | 80001.100 | 9/3/99 | 99/204,38 |
| United States | 2750-0550P | 80001.006 | 9/3/99 | 09/391,631 |
| United States | 2750-0554P | 00095.001 | 9/7/99 | 60/152,363 |
| United States | 2750-0555P | 00096.001 | 9/10/99 | 60/153,070 |
| United States | 2750-0556P | 00098.001 | 9/13/99 | 60/153,758 |
| United States | 2750-0557P | 00099.001 | 9/15/99 | 60/154,018 |
| United States | 2750-0558P | 00101.001 | 9/16/99 | 60/154,039 |
| United States | 2750-0559P | 00102.001 | 9/20/99 | 60/154,779 |
| United States | 2750-0560P | 00103.001 | 9/22/99 | 60/155,139 |
| United States | 2750-0561P | 00104.001 | 9/23/99 | 60/155,486 |
| United States | 2750-0562P | 00105.001 | 9/24/99 | 60/155,659 |
| United States | 2750-0563P | 00106.001 | 9/28/99 | 60/156,458 |
| United States | 2750-0564P | 00107.001 | 9/29/99 | 60/156,596 |
| United States | 2750-0570P | 00108.001 | 10/4/99 | 60/157,117 |
| International | 2750-0567F(PC) | 80010.100 | 10/5/99 | 99/228,55 |
| United States | 2750-0571P | 00109.001 | 10/5/99 | 60/157,753 |
| United States | 2750-0565P | 80010.002 | 10/5/99 | 09/413,198 |
| International | 2750-0568F(PC) | 80010.101 | 10/5/99 | 99/228,54 |
| United States | 2750-0566P | 80010.003 | 10/5/99 | 09/412,922 |
| International | 2750-0569F(PC) | 80010.102 | 10/5/99 | 99/228,53 |
| United States | 2750-0572P | 00110.001 | 10/6/99 | 60/157,865 |
| United States | 2750-0575P | 00111.001 | 10/7/99 | 60/158,029 |
| United States | 2750-0576P | 00112.001 | 10/8/99 | 60/158,232 |
| United States | 2750-0577P | 00113.001 | 10/12/99 | 60/158,369 |
| United States | 2750-0574P | 80145.002 | 10/13/99 | 60/159,295 |
| United States | 2750-0583P | 80148.002 | 10/13/99 | 60/159,294 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0579P | 80146.002 | 10/13/99 | 60/159,293 |
| United States | 2750-0581P | 80147.002 | 10/14/99 | 60/159,637 |
| United States | 2750-0582P | 80148.001 | 10/14/99 | 60/159,329 |
| United States | 2750-0578P | 80146.001 | 10/14/99 | 60/159,331 |
| United States | 2750-0580P | 80147.001 | 10/14/99 | 60/159,638 |
| United States | 2750-0573P | 80145.001 | 10/14/99 | 60/159,330 |
| United States | 2750-0584P | 00116.001 | 10/18/99 | 60/159,584 |
| United States | 2750-0590P | 80150.002 | 10/21/99 | 60/160,767 |
| United States | 2750-0589P | 80150.001 | 10/21/99 | 60/160,768 |
| United States | 2750-0588P | 00119.001 | 10/21/99 | 60/160,741 |
| United States | 2750-0587P | 80149.002 | 10/21/99 | 60/160,770 |
| United States | 2750-0586P | 80149.001 | 10/21/99 | 60/160,814 |
| United States | 2750-0585P | 00118.001 | 10/21/99 | 60/160,815 |
| United States | 2750-0592P | 80151.001 | 10/22/99 | 60/160,989 |
| United States | 2750-0591P | 00120.001 | 10/22/99 | 60/160,980 |
| United States | 2750-0593P | 80151.002 | 10/22/99 | 60/160,981 |
| United States | 2750-0596P | 80152.002 | 10/25/99 | 60/161,404 |
| United States | 2750-0594P | 00121.001 | 10/25/99 | 60/161,405 |
| United States | 2750-0595P | 80152.001 | 10/25/99 | 60/161,406 |
| United States | 2750-0597P | 00122.001 | 10/26/99 | 60/161,361 |
| United States | 2750-0598P | 80153.001 | 10/26/99 | 60/161,360 |
| United States | 2750-0599P | 80153.002 | 10/26/99 | 60/161,359 |
| United States | 2750-0603P | 80154.002 | 10/28/99 | 60/161,993 |
| United States | 2750-0602P | 80154.001 | 10/28/99 | 60/161,992 |
| United States | 2750-0600P | 80026.002 | 10/28/99 | 09/428,944 |
| United States | 2750-0601P | 00123.001 | 10/28/99 | 60/161,920 |
| United States | 2750-0606P | 80155.002 | 10/29/99 | 60/162,228 |
| United States | 2750-0605P | 80155.001 | 10/29/99 | 60/162,142 |
| United States | 2750-0604P | 00124.001 | 10/29/99 | 60/162,143 |
| United States | 2750-0609P | 80156.002 | 11/1/99 | 60/162,895 |
| United States | 2750-0607P | 00125.001 | 11/1/99 | 60/162,894 |
| United States | 2750-0608P | 80156.001 | 11/1/99 | 60/162,891 |
| United States | 2750-0611P | 80157.001 | 11/2/99 | 60/163,092 |
| United States | 2750-0612P | 80157.002 | 11/2/99 | 60/163,091 |
| United States | 2750-0610P | 00126.001 | 11/2/99 | 60/163,093 |
| United States | 2750-0614P | 80158.001 | 11/3/99 | 60/163,248 |
| United States | 2750-0613P | 00127.001 | 11/3/99 | 60/163,249 |
| United States | 2750-0615P | 80158.002 | 11/3/99 | 60/163,281 |
| United States | 2750-0617P | 80159.001 | 11/4/99 | 60/163,381 |
| United States | 2750-0618P | 80159.002 | 11/4/99 | 60/163,380 |
| United States | 2750-0616P | 00128.001 | 11/4/99 | 60/163,379 |
| United States | 2750-0619P | 00129.001 | 11/8/99 | 60/164,146 |
| United States | 2750-0620P | 80160.001 | 11/8/99 | 60/164,151 |
| United States | 2750-0621P | 80160.002 | 11/8/99 | 60/164,150 |
| United States | 2750-0625P | 80162.002 | 11/9/99 | 60/164,259 |
| United States | 2750-0623P | 80161.002 | 11/9/99 | 60/164,260 |
| United States | 2750-0630P | 80164.002 | 11/10/99 | 60/164,548 |
| United States | 2750-0627P | 80163.002 | 11/10/99 | 60/164,318 |
| United States | 2750-0628P | 00131.001 | 11/10/99 | 60/164,544 |
| United States | 2750-0629P | 80164.001 | 11/10/99 | 60/164,545 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0626P | 80163.001 | 11/10/99 | 60/164,321 |
| United States | 2750-0622P | 80161.001 | 11/10/99 | 60/164,319 |
| United States | 2750-0624P | 80162.001 | 11/10/99 | 60/164,317 |
| United States | 2750-0633P | 80165.002 | 11/12/99 | 60/164,960 |
| United States | 2750-0636P | 80166.002 | 11/12/99 | 60/164,962 |
| United States | 2750-0634P | 00133.001 | 11/12/99 | 60/164,870 |
| United States | 2750-0632P | 80165.001 | 11/12/99 | 60/164,871 |
| United States | 2750-0631P | 00132.001 | 11/12/99 | 60/164,961 |
| United States | 2750-0635P | 80166.001 | 11/12/99 | 60/164,959 |
| United States | 2750-0637P | 00134.001 | 11/15/99 | 60/164,927 |
| United States | 2750-0638P | 80167.001 | 11/15/99 | 60/164,929 |
| United States | 2750-0639P | 80167.002 | 11/15/99 | 60/164,926 |
| United States | 2750-0641P | 80168.001 | 11/16/99 | 60/165,671 |
| United States | 2750-0640P | 00135.001 | 11/16/99 | 60/165,669 |
| United States | 2750-0642P | 80168.002 | 11/16/99 | 60/165,661 |
| United States | 2750-0645P | 80169.002 | 11/17/99 | 60/165,911 |
| United States | 2750-0643P | 00136.001 | 11/17/99 | 60/165,919 |
| United States | 2750-0644P | 80169.001 | 11/17/99 | 60/165,918 |
| United States | 2750-0647P | 80170.001 | 11/18/99 | 60/166,173 |
| United States | 2750-0646P | 00137.001 | 11/18/99 | 60/166,157 |
| United States | 2750-0648P | 80170.002 | 11/18/99 | 60/166,158 |
| United States | 2750-0649P | 00139.001 | 11/19/99 | 60/166,419 |
| United States | 2750-0650P | 80171.001 | 11/19/99 | 60/166,411 |
| United States | 2750-0651P | 80171.002 | 11/19/99 | 60/166,412 |
| United States | 2750-0652P | 00140.001 | 11/22/99 | 60/166,733 |
| United States | 2750-0653P | 80172.001 | 11/22/99 | 60/166,750 |
| United States | 2750-0655P | 80173.002 | 11/23/99 | 60/167,362 |
| United States | 2750-0658P | 80174.002 | 11/24/99 | 60/167,235 |
| United States | 2750-0656P | 00141.001 | 11/24/99 | 60/167,233 |
| United States | 2750-0654P | 80173.001 | 11/24/99 | 60/167,382 |
| United States | 2750-0657P | 80174.001 | 11/24/99 | 60/167,234 |
| United States | 2750-0661P | 80175.002 | 11/30/99 | 60/167,902 |
| United States | 2750-0659P | 00142.001 | 11/30/99 | 60/167,904 |
| United States | 2750-0660P | 80175.001 | 11/30/99 | 60/167,908 |
| United States | 2750-0662P | 80042.002 | 12/1/99 | 09/451,320 |
| United States | 2750-0663P | 00143.001 | 12/1/99 | 60/168,232 |
| United States | 2750-0665P | 80176.002 | 12/1/99 | 60/168,231 |
| United States | 2750-0664P | 80176.001 | 12/1/99 | 60/168,233 |
| United States | 2750-0668P | 80177.002 | 12/2/99 | 60/168,548 |
| United States | 2750-0667P | 80177.001 | 12/2/99 | 60/168,549 |
| United States | 2750-0666P | 00144.001 | 12/2/99 | 60/168,546 |
| United States | 2750-0670P | 80178.001 | 12/3/99 | 60/168,673 |
| United States | 2750-0671P | 80178.002 | 12/3/99 | 60/168,674 |
| United States | 2750-0669P | 00145.001 | 12/3/99 | 60/168,675 |
| United States | 2750-0673P | 80179.001 | 12/7/99 | 60/169,278 |
| United States | 2750-0674P | 80179.002 | 12/7/99 | 60/169,302 |
| United States | 2750-0672P | 00147.001 | 12/7/99 | 60/169,298 |
| United States | 2750-0676P | 80180.002 | 12/8/99 | 60/169,691 |
| United States | 2750-0675P | 80180.001 | 12/8/99 | 60/169,692 |
| United States | 2750-0679P | 80181.002 | 12/16/99 | 60/171,098 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0678P | 80181.001 | 12/16/99 | 60/171,114 |
| United States | 2750-0677P | 00149.001 | 12/16/99 | 60/171,107 |
| United States | 2750-0683P | 80060.002 | 1/4/00 | 09/478,081 |
| United States | 2750-0684P | 80070.002 | 1/7/00 | 09/479,221 |
| International | 2750-0686F(PC) | 80070.100 | 1/7/00 | 00/004,66 |
| United States | 2750-0685P | 80183.002 | 1/19/00 | 60/176,867 |
| United States | 2750-0688P | 80184.002 | 1/19/00 | 60/176,910 |
| United States | 2750-0681P | 80182.002 | 1/19/00 | 60/176,866 |
| United States | 2750-0689P | 00152.001 | 1/26/00 | 60/178,166 |
| United States | 2750-0682P | 80183.001 | 1/27/00 | 60/178,546 |
| United States | 2750-0687P | 80184.001 | 1/27/00 | 60/178,545 |
| United States | 2750-0690P | 00153.001 | 1/27/00 | 60/178,547 |
| United States | 2750-0691P | 80185.001 | 1/27/00 | 60/177,666 |
| United States | 2750-0680P | 80182.001 | 1/27/00 | 60/178,544 |
| United States | 2750-0692P | 00155.001 | 1/28/00 | 60/178,754 |
| United States | 2750-0693P | 80186.001 | 1/28/00 | 60/178,755 |
| United States | 2750-0695P | 00157.001 | 2/1/00 | 60/179,395 |
| United States | 2750-0696P | 80187.001 | 2/1/00 | 60/179,388 |
| United States | 2750-0698P | 80188.001 | 2/3/00 | 60/180,139 |
| United States | 2750-0697P | 00158.001 | 2/3/00 | 60/180,039 |
| United States | 2750-0694P | 80084.002 | 2/3/00 | 09/497,191 |
| United States | 2750-0700P | 80189.001 | 2/4/00 | 60/180,207 |
| United States | 2750-0699P | 00159.001 | 2/4/00 | 60/180,206 |
| United States | 2750-0701P | 00160.001 | 2/7/00 | 60/180,695 |
| United States | 2750-0702P | 80190.001 | 2/7/00 | 60/180,696 |
| United States | 2750-0703P | 00161.001 | 2/9/00 | 60/181,228 |
| United States | 2750-0704P | 80191.001 | 2/9/00 | 60/181,214 |
| United States | 2750-0705P | 00162.001 | 2/10/00 | 60/181,476 |
| United States | 2750-0706P | 80192.001 | 2/10/00 | 60/181,551 |
| United States | 2750-0707P | 00163.001 | 2/15/00 | 60/182,477 |
| United States | 2750-0708P | 80193.001 | 2/15/00 | 60/182,516 |
| United States | 2750-0712P | 00164.001 | 2/15/00 | 60/182,512 |
| United States | 2750-0713P | 80194.001 | 2/15/00 | 60/182,478 |
| United States | 2750-0714P | 00165.001 | 2/17/00 | 60/183,166 |
| United States | 2750-0715P | 80195.001 | 2/17/00 | 60/183,165 |
| United States | 2750-0716P | 00167.001 | 2/24/00 | 60/184,667 |
| United States | 2750-0717P | 80196.001 | 2/24/00 | 60/184,658 |
| United States | 2750-0719P | 00168.001 | 2/25/00 | 60/185,118 |
| United States | 2750-0718P | 91022.001 | 2/25/00 | 60/185,140 |
| United States | 2750-0720P | 80197.001 | 2/25/00 | 60/185,119 |
| Europe | 2750-0709F(EP) | 80090.103 | 2/25/00 | 00/301,439 |
| Mexico | 2750-0709F(MX) | 80090.101 | 2/25/00 | 00/001,973 |
| Canada | 2750-0709F(CA) | 80090.102 | 2/25/00 | 23/006,92 |
| United States | 2750-0709P | 80090.002 | 2/25/00 | 09/513,996 |
| United States | 2750-0721P | 91023.001 | 2/28/00 | 60/185,398 |
| United States | 2750-0722P | 00169.001 | 2/28/00 | 60/185,396 |
| United States | 2750-0723P | 80198.001 | 2/28/00 | 60/185,397 |
| United States | 2750-0724P | 91024.001 | 2/29/00 | 60/185,750 |
| United States | 2750-0726P | 80199.001 | 3/1/00 | 60/186,296 |
| United States | 2750-0725P | 00170.001 | 3/1/00 | |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0710P | 80100.002 | 3/1/00 | 09/517,537 |
| United States | 2750-0727P | 91025.001 | 3/1/00 | 60/186,277 |
| United States | 2750-0729P | 00172.001 | 3/2/00 | 60/186,386 |
| United States | 2750-0730P | 80201.001 | 3/2/00 | 60/186,387 |
| United States | 2750-0728P | 80200.001 | 3/2/00 | 60/187,178 |
| United States | 2750-0711P | 00171.001 | 3/2/00 | 60/186,390 |
| United States | 2750-0731P | 91026.001 | 3/3/00 | 60/186,670 |
| United States | 2750-0732P | 00173.001 | 3/3/00 | 60/186,748 |
| United States | 2750-0733P | 80202.001 | 3/3/00 | 60/186,669 |
| United States | 2750-0734P | 00174.001 | 3/7/00 | 60/187,378 |
| United States | 2750-0735P | 91027.001 | 3/7/00 | 60/187,379 |
| United States | 2750-0736P | 00175.001 | 3/8/00 | 60/187,896 |
| United States | 2750-0737P | 80203.001 | 3/8/00 | 60/187,888 |
| United States | 2750-0738P | 91028.001 | 3/9/00 | 60/187,985 |
| United States | 2750-0740P | 80204.001 | 3/10/00 | 60/188,186 |
| United States | 2750-0741P | 91030.001 | 3/10/00 | 60/188,174 |
| United States | 2750-0742P | 00178.001 | 3/10/00 | 60/188,185 |
| United States | 2750-0743P | 80205.001 | 3/10/00 | 60/188,175 |
| United States | 2750-0739P | 00177.001 | 3/10/00 | 60/188,187 |
| United States | 2750-0744P | 91031.001 | 3/13/00 | 60/188,687 |
| United States | 2750-0745P | 00179.001 | 3/14/00 | 60/189,080 |
| United States | 2750-0746P | 80206.001 | 3/14/00 | 60/189,052 |
| United States | 2750-0747P | 91032.001 | 3/15/00 | 60/189,460 |
| United States | 2750-0748P | 00180.001 | 3/15/00 | 60/189,461 |
| United States | 2750-0749P | 80207.001 | 3/15/00 | 60/189,462 |
| United States | 2750-0756P | 80212.001 | 3/16/00 | 60/189,959 |
| United States | 2750-0754P | 91033.001 | 3/16/00 | 60/189,958 |
| United States | 2750-0750P | 80208.001 | 3/16/00 | 60/190,120 |
| United States | 2750-0751P | 80209.001 | 3/16/00 | 60/189,947 |
| United States | 2750-0752P | 80210.001 | 3/16/00 | 60/189,948 |
| United States | 2750-0753P | 80211.001 | 3/16/00 | 60/190,121 |
| United States | 2750-0757P | 91034.001 | 3/16/00 | 60/189,965 |
| United States | 2750-0755P | 00181.001 | 3/16/00 | 60/189,953 |
| United States | 2750-0759P | 80213.001 | 3/20/00 | 60/190,070 |
| United States | 2750-0760P | 91035.001 | 3/20/00 | 60/190,060 |
| United States | 2750-0758P | 00182.001 | 3/20/00 | 60/190,069 |
| United States | 2750-0762P | 80214.001 | 3/20/00 | 60/190,089 |
| United States | 2750-0761P | 00183.001 | 3/20/00 | 60/190,545 |
| United States | 2750-0763P | 00184.001 | 3/22/00 | 60/191,084 |
| United States | 2750-0764P | 80215.001 | 3/22/00 | 60/191,097 |
| United States | 2750-0766P | 00185.001 | 3/23/00 | 60/191,543 |
| United States | 2750-0765P | 91036.001 | 3/23/00 | 60/191,549 |
| United States | 2750-0767P | 80216.001 | 3/23/00 | 60/191,545 |
| United States | 2750-0769P | 00186.001 | 3/24/00 | 60/191,823 |
| United States | 2750-0770P | 80217.001 | 3/24/00 | 60/191,825 |
| United States | 2750-0768P | 91037.001 | 3/24/00 | 60/191,826 |
| United States | 2750-0771P | 91038.001 | 3/27/00 | 60/192,420 |
| United States | 2750-0772P | 00187.001 | 3/27/00 | 60/192,421 |
| United States | 2750-0773P | 80218.001 | 3/27/00 | 60/192,308 |
| United States | 2750-0775P | 00188.001 | 3/29/00 | 60/192,940 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0776P | 80219.001 | 3/29/00 | 60/192,941 |
| United States | 2750-0774P | 91039.001 | 3/29/00 | 60/192,855 |
| United States | 2750-0778P | 00189.001 | 3/30/00 | 60/193,244 |
| United States | 2750-0779P | 80220.001 | 3/30/00 | 60/193,245 |
| United States | 2750-0777P | 91040.001 | 3/30/00 | 60/193,243 |
| United States | 2750-0782P | 80221.001 | 3/31/00 | 60/193,455 |
| United States | 2750-0780P | 91041.001 | 3/31/00 | 60/193,469 |
| United States | 2750-0781P | 00190.001 | 3/31/00 | 60/193,453 |
| United States | 2750-0787P | 80222.001 | 4/4/00 | 60/194,398 |
| United States | 2750-0786P | 00191.001 | 4/4/00 | 60/194,404 |
| United States | 2750-0785P | 91042.001 | 4/4/00 | 60/194,385 |
| United States | 2750-0789P | 91043.001 | 4/5/00 | 60/194,682 |
| United States | 2750-0790P | 00192.001 | 4/5/00 | 60/194,683 |
| United States | 2750-0791P | 80223.001 | 4/5/00 | 60/194,697 |
| United States | 2750-0792P | 91044.001 | 4/5/00 | 60/194,698 |
| Europe | 2750-0783F(EP) | 91000.101 | 4/6/00 | 00/302,919 |
| United States | 2750-0793P | 00193.001 | 4/6/00 | 60/194,874 |
| Mexico | 2750-0783F(MX) | 91000.100 | 4/6/00 | 00/003,391 |
| Canada | 2750-0783F(CA) | 91000.102 | 4/6/00 | 02/302,828 |
| United States | 2750-0783P | 91000.002 | 4/6/00 | 09/543,680 |
| United States | 2750-0784P | 91045.001 | 4/6/00 | 60/194,884 |
| United States | 2750-0795P | 00194.001 | 4/6/00 | 60/194,885 |
| United States | 2750-0794P | 80224.001 | 4/6/00 | 60/194,872 |
| United States | 2750-0796P | 80225.001 | 4/6/00 | 60/195,045 |
| United States | 2750-0799P | 80226.001 | 4/7/00 | 60/195,257 |
| United States | 2750-0797P | 91046.001 | 4/7/00 | 60/195,258 |
| United States | 2750-0798P | 00195.001 | 4/7/00 | 60/195,283 |
| United States | 2750-0802P | 91047.001 | 4/11/00 | 60/196,168 |
| United States | 2750-0804P | 80228.001 | 4/11/00 | 60/196,089 |
| United States | 2750-0801P | 80227.002 | 4/11/00 | 60/196,211 |
| United States | 2750-0803P | 00196.001 | 4/11/00 | 60/196,169 |
| United States | 2750-0800P | 80227.001 | 4/12/00 | 60/196,212 |
| United States | 2750-0805P | 91048.001 | 4/12/00 | 60/196,483 |
| United States | 2750-0806P | 00197.001 | 4/12/00 | 60/196,487 |
| United States | 2750-0807P | 80229.001 | 4/12/00 | 60/196,289 |
| United States | 2750-0808P | 00200.001 | 4/12/00 | 60/196,485 |
| United States | 2750-0809P | 80230.001 | 4/12/00 | 60/196,486 |
| United States | 2750-0811P | 80231.002 | 4/13/00 | 60/196,213 |
| United States | 2750-0810P | 80231.001 | 4/14/00 | |
| United States | 2750-0814P | 91049.001 | 4/14/00 | 60/197,397 |
| United States | 2750-0815P | 00201.001 | 4/17/00 | 60/197,687 |
| United States | 2750-0813P | 80232.002 | 4/17/00 | 60/197,871 |
| United States | 2750-0819P | 80234.001 | 4/17/00 | 60/197,671 |
| United States | 2750-0816P | 80233.001 | 4/17/00 | 60/197,678 |
| United States | 2750-0818P | 00202.001 | 4/17/00 | 60/198,133 |
| United States | 2750-0812P | 80232.001 | 4/17/00 | 60/197,870 |
| United States | 2750-0817P | 91050.001 | 4/17/00 | 60/198,268 |
| United States | 2750-0820P | 91051.001 | 4/19/00 | 60/198,400 |
| United States | 2750-0821P | 00203.001 | 4/19/00 | 60/198,386 |
| United States | 2750-0822P | 80235.001 | 4/19/00 | 60/198,373 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0823P | 91052.001 | 4/20/00 | 60/198,629 |
| United States | 2750-0824P | 00204.001 | 4/20/00 | 60/198,619 |
| United States | 2750-0825P | 80236.001 | 4/20/00 | 60/198,623 |
| United States | 2750-0827P | 00206.001 | 4/21/00 | 60/198,767 |
| United States | 2750-0828P | 80237.001 | 4/21/00 | 60/198,763 |
| United States | 2750-0826P | 91053.001 | 4/21/00 | 60/198,765 |
| United States | 2750-0829P | 91054.001 | 4/24/00 | 60/199,123 |
| United States | 2750-0831P | 80238.001 | 4/24/00 | 60/199,122 |
| United States | 2750-0830P | 00207.001 | 4/24/00 | 60/199,124 |
| United States | 2750-0832P | 92001.001 | 4/26/00 | 60/200,034 |
| United States | 2750-0833P | 92002.001 | 4/26/00 | 60/200,031 |
| United States | 2750-0834P | 00208.001 | 4/26/00 | 60/199,828 |
| United States | 2750-0835P | 80239.001 | 4/26/00 | 60/199,818 |
| United States | 2750-0836P | 00210.001 | 4/27/00 | 60/200,103 |
| United States | 2750-0837P | 80240.001 | 4/27/00 | 60/200,102 |
| United States | 2750-0788P | 80123.002 | 4/28/00 | 09/559,232 |
| United States | 2750-0844P | 80242.002 | 4/28/00 | 60/200,373 |
| United States | 2750-0846P | 80243.002 | 4/28/00 | 60/200,773 |
| United States | 2750-0848P | 80244.002 | 5/1/00 | 60/200,761 |
| United States | 2750-0841P | 92001.002 | 5/1/00 | 60/201,017 |
| United States | 2750-0842P | 92002.002 | 5/1/00 | 60/201,018 |
| United States | 2750-0847P | 80244.001 | 5/1/00 | 60/200,762 |
| United States | 2750-0845P | 80243.001 | 5/1/00 | 60/201,016 |
| United States | 2750-0843P | 80242.001 | 5/1/00 | 60/200,763 |
| United States | 2750-0839P | 80241.001 | 5/1/00 | 60/200,879 |
| United States | 2750-0840P | 91055.001 | 5/1/00 | 60/200,885 |
| United States | 2750-0838P | 00211.001 | 5/2/00 | 60/201,275 |
| United States | 2750-0849P | 91056.001 | 5/2/00 | 60/201,279 |
| United States | 2750-0850P | 80245.001 | 5/2/00 | 60/201,305 |
| United States | 2750-0856P | 91057.001 | 5/4/00 | 60/201,751 |
| Mexico | 2750-0851F(MX) | 91002.102 | 5/4/00 | 00/004,406 |
| United States | 2750-0857P | 00212.001 | 5/4/00 | 60/201,740 |
| United States | 2750-0858P | 80246.001 | 5/4/00 | 60/201,750 |
| United States | 2750-0852P | 80126.002 | 5/4/00 | 09/566,262 |
| Canada | 2750-0851F(CA) | 91002.100 | 5/5/00 | 02/305,695 |
| United States | 2750-0859P | 91058.001 | 5/5/00 | 60/202,178 |
| United States | 2750-0855P | 80130.002 | 5/5/00 | 09/565,310 |
| United States | 2750-0854P | 80129.002 | 5/5/00 | 09/565,307 |
| United States | 2750-0853P | 80127.002 | 5/5/00 | 09/565,309 |
| United States | 2750-0860P | 00213.001 | 5/5/00 | 60/202,112 |
| United States | 2750-0861P | 80247.001 | 5/5/00 | 60/202,180 |
| Europe | 2750-0851F(EP) | 91002.101 | 5/5/00 | 00/303,770 |
| United States | 2750-0851P | 91002.002 | 5/5/00 | 09/565,308 |
| United States | 2750-0862P | 00214.001 | 5/9/00 | 60/202,914 |
| United States | 2750-0866P | 80249.001 | 5/9/00 | 60/202,634 |
| United States | 2750-0865P | 00215.001 | 5/9/00 | 60/202,919 |
| United States | 2750-0863P | 80248.001 | 5/9/00 | 60/202,636 |
| United States | 2750-0864P | 91059.001 | 5/9/00 | 60/202,915 |
| United States | 2750-0878P | 00216.001 | 5/10/00 | 60/202,968 |
| United States | 2750-0877P | 91060.001 | 5/10/00 | 60/202,969 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0879P | 80252.001 | 5/10/00 | 60/202,963 |
| United States | 2750-0871P | 80131.002 | 5/11/00 | 09/572,408 |
| United States | 2750-0881P | 00217.001 | 5/11/00 | 60/203,457 |
| United States | 2750-0868P | 80250.002 | 5/11/00 | 60/203,672 |
| United States | 2750-0867P | 80250.001 | 5/11/00 | 60/203,671 |
| United States | 2750-0882P | 80253.001 | 5/11/00 | 60/203,279 |
| United States | 2750-0870P | 80251.002 | 5/11/00 | 60/203,622 |
| United States | 2750-0880P | 91061.001 | 5/11/00 | 60/203,458 |
| United States | 2750-0869P | 80251.001 | 5/11/00 | 60/203,669 |
| United States | 2750-0873P | 00025.002 | 5/12/00 | 09/570,582 |
| Europe | 2750-0875F(EP) | 91006.101 | 5/12/00 | 00/304,017 |
| Mexico | 2750-0875F(MX) | 91006.102 | 5/12/00 | 00/004,628 |
| United States | 2750-0874P | 80116.002 | 5/12/00 | 09/570,738 |
| Canada | 2750-0875F(CA) | 91006.100 | 5/12/00 | 02/306,232 |
| United States | 2750-0875P | 91006.002 | 5/12/00 | 09/570,581 |
| United States | 2750-0883P | 91062.001 | 5/12/00 | 60/203,911 |
| United States | 2750-0872P | 80117.002 | 5/12/00 | 09/570,768 |
| United States | 2750-0884P | 00219.001 | 5/12/00 | 60/203,916 |
| United States | 2750-0885P | 80254.001 | 5/12/00 | 60/203,915 |
| United States | 2750-0887P | 00220.001 | 5/15/00 | 60/204,388 |
| United States | 2750-0886P | 91063.001 | 5/15/00 | 60/204,395 |
| United States | 2750-0888P | 80255.001 | 5/15/00 | 60/204,122 |
| United States | 2750-0891P | 00221.001 | 5/16/00 | 60/204,568 |
| United States | 2750-0892P | 80256.001 | 5/16/00 | 60/204,569 |
| United States | 2750-0889P | 92001.003 | 5/17/00 | 60/205,233 |
| United States | 2750-0894P | 80257.001 | 5/17/00 | 60/204,829 |
| United States | 2750-0893P | 00222.001 | 5/17/00 | 60/204,830 |
| United States | 2750-0890P | 92002.003 | 5/17/00 | 60/205,325 |
| Mexico | 2750-0876F(MX) | 91007.102 | 5/18/00 | 00/004,850 |
| Canada | 2750-0876F(CA) | 91007.100 | 5/18/00 | 02/306,202 |
| United States | 2750-0876P | 91007.002 | 5/18/00 | 09/573,655 |
| Europe | 2750-0876F(EP) | 91007.101 | 5/18/00 | 00/304,161 |
| United States | 2750-0896P | 80258.001 | 5/18/00 | 60/205,058 |
| United States | 2750-0895P | 00223.001 | 5/18/00 | 60/205,201 |
| United States | 2750-0897P | 00224.001 | 5/19/00 | 60/205,242 |
| United States | 2750-0898P | 80259.001 | 5/19/00 | 60/205,243 |
| United States | 2750-0899P | 91064.001 | 5/22/00 | 60/205,574 |
| United States | 2750-0901P | 80260.001 | 5/22/00 | 60/205,576 |
| United States | 2750-0900P | 00225.001 | 5/22/00 | 60/205,572 |
| United States | 2750-0903P | 80261.001 | 5/23/00 | 60/206,319 |
| United States | 2750-0902P | 00226.001 | 5/23/00 | 60/206,316 |
| United States | 2750-0905P | 80262.001 | 5/24/00 | 60/206,545 |
| United States | 2750-0904P | 00227.001 | 5/24/00 | 60/206,553 |
| United States | 2750-0906P | 91065.001 | 5/25/00 | 60/206,988 |
| United States | 2750-0907P | 00228.001 | 5/26/00 | 60/207,367 |
| United States | 2750-0910P | 00229.001 | 5/26/00 | 60/207,239 |
| United States | 2750-0909P | 91066.001 | 5/26/00 | 60/207,242 |
| United States | 2750-0911P | 80264.001 | 5/26/00 | 60/207,354 |
| United States | 2750-0908P | 80263.001 | 5/26/00 | 60/207,243 |
| United States | 2750-0912P | 91067.001 | 5/30/00 | 60/207,291 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0914P | 80265.001 | 5/30/00 | 60/207,329 |
| United States | 2750-0913P | 00230.001 | 5/30/00 | 60/207,452 |
| United States | 2750-0919P | 91068.001 | 6/1/00 | 60/208,324 |
| United States | 2750-0921P | 80268.001 | 6/1/00 | 60/208,312 |
| United States | 2750-0920P | 00231.001 | 6/1/00 | 60/208,329 |
| United States | 2750-0918P | 80267.002 | 6/1/00 | 60/208,648 |
| United States | 2750-0916P | 80266.002 | 6/1/00 | 60/208,421 |
| United States | 2750-0917P | 80267.001 | 6/2/00 | 60/208,649 |
| United States | 2750-0915P | 80266.001 | 6/2/00 | 60/209,338 |
| United States | 2750-0926P | 00233.001 | 6/5/00 | 60/208,921 |
| United States | 2750-0927P | 80270.001 | 6/5/00 | 60/208,920 |
| United States | 2750-0924P | 80269.001 | 6/5/00 | 60/208,918 |
| United States | 2750-0923P | 00232.001 | 6/5/00 | 60/208,910 |
| United States | 2750-0922P | 91069.001 | 6/5/00 | 60/208,919 |
| United States | 2750-0925P | 91070.001 | 6/5/00 | 60/208,917 |
| United States | 2750-0931P | 80271.001 | 6/8/00 | 60/210,006 |
| United States | 2750-0930P | 00234.001 | 6/8/00 | 60/210,012 |
| United States | 2750-0929P | 91071.001 | 6/8/00 | 60/210,008 |
| Mexico | 2750-0928F(MX) | 00033.102 | 6/9/00 | 00/005,740 |
| Mexico | 2750-1037F(MX) | | 6/9/00 | 00/005,741 |
| Canada | 2750-0928F(CA) | 00033.100 | 6/9/00 | |
| United States | 2750-0928P | 00033.003 | 6/9/00 | 09/592,459 |
| United States | 2750-0932P | 00235.001 | 6/9/00 | 60/210,670 |
| United States | 2750-0933P | 80272.001 | 6/9/00 | 60/210,564 |
| Europe | 2750-0928F(EP) | 00033.101 | 6/12/00 | 00/304,943 |
| United States | 2750-0935P | 00237.001 | 6/13/00 | 60/211,213 |
| United States | 2750-0936P | 80273.001 | 6/13/00 | 60/211,214 |
| United States | 2750-0937P | 91072.001 | 6/13/00 | 60/211,210 |
| Europe | 2750-0934F(EP) | 00034.101 | 6/14/00 | 00/305,026 |
| United States | 2750-0934P | 00034.002 | 6/14/00 | 09/593,710 |
| Mexico | 2750-0934F(MX) | 00034.102 | 6/14/00 | 00/005,842 |
| Canada | 2750-0934F(CA) | 00034.100 | 6/14/00 | |
| United States | 2750-0939P | 80274.001 | 6/15/00 | 60/211,540 |
| United States | 2750-0940P | 91074.001 | 6/15/00 | 60/211,538 |
| United States | 2750-0938P | 00238.001 | 6/15/00 | 60/211,539 |
| Mexico | 2750-0942F(MX) | 00038.102 | 6/16/00 | |
| Mexico | 2750-0943F(MX) | 00039.102 | 6/16/00 | |
| Mexico | 2750-0941F(MX) | 00037.102 | 6/16/00 | 00/005,950 |
| Europe | 2750-0941F(EP) | 00037.101 | 6/16/00 | 00/305,144 |
| Canada | 2750-0941F(CA) | 00037.100 | 6/16/00 | |
| United States | 2750-0945P | 80132.014 | 6/16/00 | 09/595,333 |
| Europe | 2750-0943F(EP) | 00039.101 | 6/16/00 | |
| United States | 2750-0943P | 00039.002 | 6/16/00 | 09/596,577 |
| Canada | 2750-0943F(CA) | 00039.100 | 6/16/00 | |
| United States | 2750-0954P | 80132.023 | 6/16/00 | 09/594,599 |
| United States | 2750-0947P | 80132.016 | 6/16/00 | 09/595,335 |
| Europe | 2750-0942F(EP) | 00038.101 | 6/16/00 | |
| United States | 2750-0951P | 80132.020 | 6/16/00 | 09/594,595 |
| Canada | 2750-0942F(CA) | 00038.100 | 6/16/00 | |
| United States | 2750-0949P | 80132.018 | 6/16/00 | 09/595,332 |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0952P | 80132.021 | 6/16/00 | 09/594,597 |
| United States | 2750-0955P | 80132.024 | 6/16/00 | 09/595,331 |
| United States | 2750-0942P | 00038.002 | 6/16/00 | 09/595,326 |
| United States | 2750-0946P | 80132.015 | 6/16/00 | 09/595,328 |
| United States | 2750-0948P | 80132.017 | 6/16/00 | 09/595,329 |
| United States | 2750-0944P | 80132.013 | 6/16/00 | 09/595,330 |
| United States | 2750-0941P | 00037.002 | 6/16/00 | 09/595,334 |
| United States | 2750-0950P | 80132.019 | 6/16/00 | 09/594,598 |
| United States | 2750-0953P | 80132.022 | 6/16/00 | 09/595,298 |
| United States | 2750-0958P | 91075.001 | 6/19/00 | 60/212,623 |
| United States | 2750-0957P | 80275.001 | 6/19/00 | 60/212,649 |
| United States | 2750-0956P | 00239.001 | 6/19/00 | 60/212,414 |
| United States | 2750-0959P | 00240.001 | 6/20/00 | 60/212,677 |
| United States | 2750-0960P | 80276.001 | 6/20/00 | 60/212,713 |
| United States | 2750-0961P | 91076.001 | 6/20/00 | 60/212,727 |
| Mexico | 2750-0971F(MX) | 00042.102 | 6/21/00 | 00/006,142 |
| United States | 2750-0971P | 00042.003 | 6/21/00 | 09/602,660 |
| Canada | 2750-0971F(CA) | 00042.100 | 6/21/00 | 02/309,874 |
| Europe | 2750-0971F(EP) | 00042.101 | 6/21/00 | 00/305,249 |
| United States | 2750-0967P | 91079.001 | 6/22/00 | 60/213,270 |
| United States | 2750-0965P | 00246.001 | 6/22/00 | 60/213,221 |
| Mexico | 2750-0972F(MX) | 00043.102 | 6/22/00 | 00/006,625 |
| Europe | 2750-0972F(EP) | 00043.101 | 6/22/00 | 00/305,270 |
| Canada | 2750-0972F(CA) | 00043.100 | 6/22/00 | 02/309,793 |
| United States | 2750-0972P | 00043.002 | 6/22/00 | 09/602,152 |
| United States | 2750-0964P | 91077.001 | 6/22/00 | 60/213,964 |
| United States | 2750-0963P | 80277.001 | 6/22/00 | 60/213,249 |
| United States | 2750-0962P | 00242.001 | 6/22/00 | 60/213,271 |
| United States | 2750-0966P | 80278.001 | 6/22/00 | 60/213,220 |
| Canada | 2750-0974F(CA) | 00042.100 | 6/23/00 | |
| Mexico | 2750-0973F(MX) | 00044.102 | 6/23/00 | 00/006,267 |
| Europe | 2750-0974F(EP) | 00042.101 | 6/23/00 | |
| Mexico | 2750-0974F(MX) | 00042.102 | 6/23/00 | |
| Mexico | 2750-0975F(MX) | 00045.102 | 6/23/00 | |
| Europe | 2750-0975F(EP) | 00045.101 | 6/23/00 | |
| United States | 2750-0973P | 00044.002 | 6/23/00 | 09/602,205 |
| United States | 2750-0975P | 00045.002 | 6/23/00 | |
| Canada | 2750-0973F(CA) | 00044.100 | 6/23/00 | 02/309,889 |
| Europe | 2750-0973F(EP) | 00044.101 | 6/23/00 | 00/305,305 |
| Canada | 2750-0975F(CA) | 00045.100 | 6/23/00 | |
| United States | 2750-0968P | 00247.001 | 6/27/00 | |
| United States | 2750-1035P | 00248.001 | 6/27/00 | |
| United States | 2750-0969P | 80279.001 | 6/27/00 | |
| United States | 2750-1036P | 80280.001 | 6/27/00 | |
| United States | 2750-0970P | 91080.001 | 6/27/00 | 60/214,524 |
| United States | 2750-1039P | 80281.001 | 6/28/00 | |
| Europe | 2750-0976F(EP) | 00046.101 | 6/28/00 | |
| United States | 2750-1038P | 00249.001 | 6/28/00 | |
| Mexico | 2750-0976F(MX) | 00046.102 | 6/28/00 | |
| Canada | 2750-0976F(CA) | 00046.100 | 6/28/00 | |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0976P | 00046.002 | 6/28/00 | 09/605,843 |
| United States | 2750-0977P | 00048.002 | 6/29/00 | |
| Mexico | 2750-0977F(MX) | 00048.102 | 6/29/00 | |
| Canada | 2750-0977F(CA) | 00048.100 | 6/29/00 | |
| Europe | 2750-0977F(EP) | 00048.101 | 6/29/00 | |
| United States | 2750-0979P | 00050.002 | 6/30/00 | 09/607,081 |
| Mexico | 2750-0980F(MX) | 00051.102 | 6/30/00 | |
| United States | 2750-1040P | 00250.001 | 6/30/00 | |
| Europe | 2750-0981F(EP) | 00052.101 | 6/30/00 | |
| Mexico | 2750-0978F(MX) | 00049.102 | 6/30/00 | |
| Europe | 2750-0978F(EP) | 00049.101 | 6/30/00 | |
| Canada | 2750-0978F(CA) | 00049.100 | 6/30/00 | |
| Mexico | 2750-0979F(MX) | 00050.102 | 6/30/00 | |
| Canada | 2750-0979F(CA) | 00050.100 | 6/30/00 | |
| United States | 2750-0978P | 00049.002 | 6/30/00 | |
| United States | 2750-1041P | 80282.001 | 6/30/00 | |
| Mexico | 2750-0981F(MX) | 00052.102 | 6/30/00 | |
| Canada | 2750-0981F(CA) | 00052.100 | 6/30/00 | |
| United States | 2750-0981P | 00052.002 | 6/30/00 | |
| United States | 2750-0980P | 00051.002 | 6/30/00 | |
| Canada | 2750-0980F(CA) | 00051.100 | 6/30/00 | |
| Europe | 2750-0980F(EP) | 00051.101 | 6/30/00 | |
| Europe | 2750-0979F(EP) | 00050.101 | 6/30/00 | |
| United States | 2750-1042P | 00252.001 | 7/5/00 | |
| United States | 2750-1043P | 80283.001 | 7/5/00 | |
| Europe | 2750-0982F(EP) | 00053.101 | 7/6/00 | |
| Mexico | 2750-0982F(MX) | 00053.102 | 7/6/00 | |
| Canada | 2750-0982F(CA) | 00053.100 | 7/6/00 | |
| United States | 2750-0982P | 00053.002 | 7/6/00 | |
| Europe | 2750-0983F(EP) | 00054.101 | 7/7/00 | |
| United States | 2750-0983P | 00054.002 | 7/7/00 | |
| Mexico | 2750-0984F(MX) | 00058.102 | 7/7/00 | |
| Europe | 2750-0984F(EP) | 00058.101 | 7/7/00 | |
| Canada | 2750-0984F(CA) | 00058.100 | 7/7/00 | |
| Mexico | 2750-0983F(MX) | 00054.102 | 7/7/00 | |
| Canada | 2750-0983F(CA) | 00054.100 | 7/7/00 | |
| United States | 2750-0984P | 00058.002 | 7/7/00 | |
| United States | 2750-1045P | 00253.001 | 7/11/00 | |
| United States | 2750-1046P | 80284.001 | 7/11/00 | |
| United States | 2750-1044P | 91081.001 | 7/11/00 | |
| Canada | 2750-0985F(CA) | 00059.100 | 7/12/00 | |
| United States | 2750-0985P | 00059.002 | 7/12/00 | 09/615,007 |
| Mexico | 2750-0985F(MX) | 00059.102 | 7/12/00 | |
| United States | 2750-1052P | 80287.002 | 7/12/00 | |
| Europe | 2750-0985F(EP) | 00059.101 | 7/12/00 | |
| United States | 2750-1050P | 80286.002 | 7/12/00 | |
| Canada | 2750-0986F(CA) | 00060.100 | 7/13/00 | |
| United States | 2750-1054P | 80288.002 | 7/13/00 | |
| Mexico | 2750-0986F(MX) | 00060.102 | 7/13/00 | |
| Europe | 2750-0986F(EP) | 00060.101 | 7/13/00 | |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-0986P | 00060.002 | 7/13/00 | |
| United States | 2750-1061P | 80134.018 | 7/14/00 | |
| United States | 2750-1060P | 80134.017 | 7/14/00 | |
| United States | 2750-0987P | 00061.002 | 7/14/00 | |
| Mexico | 2750-0988F(MX) | 00062.102 | 7/14/00 | |
| Europe | 2750-0987F(EP) | 00061.101 | 7/14/00 | |
| Mexico | 2750-0987F(MX) | 00061.102 | 7/14/00 | |
| United States | 2750-0988P | 00062.002 | 7/14/00 | |
| Europe | 2750-0988F(EP) | 00062.101 | 7/14/00 | |
| Canada | 2750-0987F(CA) | 00061.100 | 7/14/00 | |
| United States | 2750-1048P | 80285.002 | 7/14/00 | |
| Canada | 2750-0988F(CA) | 00062.100 | 7/14/00 | |
| United States | 2750-1057P | 80291.001 | 7/18/00 | |
| United States | 2750-1056P | 00254.001 | 7/18/00 | |
| United States | 2750-1055P | 91082.001 | 7/18/00 | |
| United States | 2750-0989P | 00064.002 | 7/19/00 | 09/620,421 |
| Canada | 2750-0989F(CA) | 00064.100 | 7/19/00 | |
| Europe | 2750-0989F(EP) | 00064.101 | 7/19/00 | |
| United States | 2750-1062P | 80134.020 | 7/19/00 | |
| United States | 2750-1064P | 80134.024 | 7/19/00 | |
| United States | 2750-1063P | 80134.022 | 7/19/00 | |
| Mexico | 2750-0989F(MX) | 00064.102 | 7/19/00 | |
| United States | 2750-0990P | 00065.002 | 7/20/00 | |
| Canada | 2750-0990F(CA) | 00065.100 | 7/20/00 | |
| Europe | 2750-0990F(EP) | 00065.101 | 7/20/00 | |
| United States | 2750-1065P | 80134.026 | 7/20/00 | |
| Mexico | 2750-0990F(MX) | 00065.102 | 7/20/00 | |
| United States | 2750-1066P | 80135.004 | 7/20/00 | |
| United States | 2750-0993P | 00069.002 | 7/21/00 | |
| United States | 2750-1073P | 80134.025 | 7/21/00 | |
| United States | 2750-1072P | 80135.003 | 7/21/00 | |
| United States | 2750-0992P | 00067.002 | 7/21/00 | |
| Canada | 2750-0992F(CA) | 00067.100 | 7/21/00 | |
| Europe | 2750-0992F(EP) | 00067.101 | 7/21/00 | |
| United States | 2750-1070P | 80134.019 | 7/21/00 | |
| Mexico | 2750-0993F(MX) | 00069.102 | 7/21/00 | |
| United States | 2750-1071P | 80134.021 | 7/21/00 | |
| Canada | 2750-0993F(CA) | 00069.100 | 7/21/00 | |
| United States | 2750-1067P | 80134.015 | 7/21/00 | |
| United States | 2750-1069P | 80134.023 | 7/21/00 | |
| Mexico | 2750-0992F(MX) | 00067.102 | 7/21/00 | |
| United States | 2750-1068P | 80134.016 | 7/21/00 | |
| United States | 2750-0991P | 00066.002 | 7/21/00 | 09/621,630 |
| Canada | 2750-0991F(CA) | 00066.100 | 7/21/00 | |
| Europe | 2750-0991F(EP) | 00066.101 | 7/21/00 | |
| Mexico | 2750-0991F(MX) | 00066.102 | 7/21/00 | |
| Europe | 2750-0993F(EP) | 00069.101 | 7/21/00 | |
| United States | 2750-1058P | 91083.001 | 7/25/00 | |
| United States | 2750-1081P | 80293.001 | 7/25/00 | |
| United States | 2750-1080P | 00256.001 | 7/25/00 | |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-1079P | 80292.001 | 7/25/00 | |
| United States | 2750-1059P | 00255.001 | 7/25/00 | |
| United States | 2750-0994P | 00070.002 | 7/26/00 | |
| Mexico | 2750-0994F(MX) | 00070.102 | 7/26/00 | |
| Europe | 2750-0994F(EP) | 00070.101 | 7/26/00 | |
| Canada | 2750-0994F(CA) | 00070.100 | 7/26/00 | |
| United States | 2750-0995P | 00071.002 | 7/27/00 | |
| Canada | 2750-0995F(CA) | 00071.100 | 7/27/00 | |
| Europe | 2750-0995F(EP) | 00071.101 | 7/27/00 | |
| United States | 2750-1075P | 80136.005 | 7/27/00 | |
| Mexico | 2750-0995F(MX) | 00071.102 | 7/27/00 | |
| United States | 2750-1074P | 80136.004 | 7/27/00 | |
| Mexico | 2750-0996F(MX) | 00072.102 | 7/28/00 | |
| United States | 2750-0996P | 00072.002 | 7/28/00 | |
| Canada | 2750-0996F(CA) | 00072.100 | 7/28/00 | |
| Europe | 2750-0996F(EP) | 00072.101 | 7/28/00 | |
| United States | 2750-1049P | 80286.001 | 8/1/00 | |
| Europe | 2750-0997F(EP) | 00073.101 | 8/2/00 | |
| Canada | 2750-0997F(CA) | 00073.100 | 8/2/00 | |
| United States | 2750-0997P | 00073.002 | 8/2/00 | |
| United States | 2750-1077P | 80137.004 | 8/2/00 | |
| Mexico | 2750-0997F(MX) | 00073.102 | 8/2/00 | |
| United States | 2750-1076P | 80137.003 | 8/2/00 | |
| Mexico | 2750-0998F(MX) | 00074.102 | 8/3/00 | |
| Europe | 2750-0998F(EP) | 00074.101 | 8/3/00 | |
| United States | 2750-1053P | 80288.002 | 8/3/00 | |
| United States | 2750-0998P | 00074.002 | 8/3/00 | 09/632,349 |
| United States | 2750-1051P | 80287.001 | 8/3/00 | |
| Canada | 2750-0998F(CA) | 00074.100 | 8/3/00 | |
| Europe | 2750-1001F(EP) | 00079.101 | 8/4/00 | |
| Mexico | 2750-1000F(MX) | 00077.102 | 8/4/00 | |
| Mexico | 2750-1001F(MX) | 00079.102 | 8/4/00 | |
| Canada | 2750-1001F(CA) | 00079.100 | 8/4/00 | |
| United States | 2750-1000P | 00077.002 | 8/4/00 | |
| Canada | 2750-1000F(CA) | 00077.100 | 8/4/00 | |
| Europe | 2750-1000F(EP) | 00077.101 | 8/4/00 | |
| United States | 2750-1078P | 80138.003 | 8/4/00 | |
| Europe | 2750-0999F(EP) | 00076.101 | 8/4/00 | |
| United States | 2750-1001P | 00079.002 | 8/4/00 | |
| Canada | 2750-0999F(CA) | 00076.100 | 8/4/00 | |
| United States | 2750-1092P | 80138.004 | 8/4/00 | |
| United States | 2750-0999P | 00076.002 | 8/4/00 | |
| Mexico | 2750-0999F(MX) | 00076.102 | 8/4/00 | |
| United States | 2750-1047P | 80285.001 | 8/7/00 | |
| Europe | 2750-1002F(EP) | 00080.101 | 8/9/00 | |
| United States | 2750-1094P | 80139.004 | 8/9/00 | |
| Mexico | 2750-1002F(MX) | 00080.102 | 8/9/00 | |
| Canada | 2750-1002F(CA) | 00080.100 | 8/9/00 | |
| United States | 2750-1002P | 00080.002 | 8/9/00 | |
| United States | 2750-1115P | 92002.004 | 8/9/00 | |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-1114P | 92001.004 | 8/9/00 | |
| United States | 2750-1093P | 80139.003 | 8/10/00 | |
| Europe | 2750-1003F(EP) | 00081.101 | 8/11/00 | |
| Mexico | 2750-1004F(MX) | 00082.102 | 8/11/00 | |
| Canada | 2750-1004F(CA) | 00082.100 | 8/11/00 | |
| United States | 2750-1004P | 00082.002 | 8/11/00 | |
| Mexico | 2750-1005F(MX) | 00083.102 | 8/11/00 | |
| Europe | 2750-1004F(EP) | 00082.101 | 8/11/00 | |
| Europe | 2750-1005F(EP) | 00083.101 | 8/11/00 | |
| United States | 2750-1102P | 80141.006 | 8/11/00 | |
| Mexico | 2750-1003F(MX) | 00081.102 | 8/11/00 | |
| Canada | 2750-1003F(CA) | 00081.100 | 8/11/00 | |
| United States | 2750-1003P | 00081.002 | 8/11/00 | |
| United States | 2750-1103P | 80141.007 | 8/11/00 | |
| United States | 2750-1101P | 80141.005 | 8/11/00 | |
| Canada | 2750-1005F(CA) | 00083.100 | 8/11/00 | |
| United States | 2750-1005P | 00083.002 | 8/11/00 | |
| United States | 2750-1096P | 80142.004 | 8/11/00 | |
| United States | 2750-1104P | 80141.008 | 8/11/00 | |
| United States | 2750-1082P | 80298.001 | 8/14/00 | |
| United States | 2750-1083P | 91084.001 | 8/14/00 | |
| United States | 2750-1084P | 80300.001 | 8/15/00 | |
| United States | 2750-1085P | 91085.001 | 8/15/00 | |
| United States | 2750-1106P | 80294.002 | 8/16/00 | |
| United States | 2750-1108P | 80295.002 | 8/16/00 | |
| United States | 2750-1095P | 80142.003 | 8/16/00 | |
| Mexico | 2750-1006F(MX) | 00084.102 | 8/17/00 | |
| Canada | 2750-1006F(CA) | 00084.100 | 8/17/00 | |
| Europe | 2750-1006F(EP) | 00084.101 | 8/17/00 | |
| United States | 2750-1006P | 00084.002 | 8/17/00 | |
| United States | 2750-1109P | 80296.001 | 8/18/00 | |
| Canada | 2750-1009F(CA) | 00087.100 | 8/18/00 | |
| United States | 2750-1009P | 00087.002 | 8/18/00 | |
| Europe | 2750-1009F(EP) | 00087.101 | 8/18/00 | |
| United States | 2750-1098P | 80143.004 | 8/18/00 | |
| mexico | 2750-1009F(MX) | 00087.102 | 8/18/00 | |
| United States | 2750-1105P | 80294.001 | 8/18/00 | |
| United States | 2750-1107P | 80295.001 | 8/18/00 | |
| United States | 2750-1086P | 80301.001 | 8/21/00 | |
| United States | 2750-1087P | 91086.001 | 8/21/00 | |
| United States | 2750-1162P | 80307.001 | 8/23/00 | |
| United States | 2750-1163P | 91087.001 | 8/23/00 | |
| United States | 2750-1097P | 80143.003 | 8/23/00 | |
| United States | 2750-1010P | 00088.002 | 8/23/00 | |
| Canada | 2750-1010F(CA) | 00088.100 | 8/23/00 | |
| Europe | 2750-1010F(EP) | 00088.101 | 8/23/00 | |
| Mexico | 2750-1010F(MX) | 00088.102 | 8/23/00 | |
| United States | 2750-1152P | 80289.037 | 8/25/00 | |
| United States | 2750-1151P | 80289.036 | 8/25/00 | |
| United States | 2750-1100P | 80144.004 | 8/25/00 | |

| Country | Attorney No. | Client No. | Filed | Application No. |
|---|---|---|---|---|
| United States | 2750-1099P | 80144.003 | 8/25/00 | |
| Europe | 2750-1014F(EP) | 00092.101 | 8/30/00 | |
| Canada | 2750-1014F(CA) | 00092.100 | 8/30/00 | |
| United States | 2750-1014P | 00092.002 | 8/30/00 | |
| Mexico | 2750-1014F(MX) | 00092.102 | 8/30/00 | |
| Canada | 2750-1015F(CA) | 00093.100 | 8/31/00 | |
| Europe | 2750-1015F(EP) | 00093.101 | 8/31/00 | |
| Mexico | 2750-1015F(MX) | 00093.102 | 8/31/00 | |
| United States | 2750-1015P | 00093.002 | 8/31/00 | |
| Canada | 2750-1016F(CA) | 00094.100 | 9/1/00 | |
| Europe | 2750-1016F(EP) | 00094.101 | 9/1/00 | |
| Mexico | 2750-1016F(MX) | 00094.102 | 9/1/00 | |
| United States | 2750-1016P | 00094.002 | 9/1/00 | |
| United States | 2750-1227P | 80309.001 | 9/6/00 | |
| United States | 2750-1228P | 91089.001 | 9/6/00 | |
| Canada | 2750-1017F(CA) | 00095.100 | 9/7/00 | |
| Mexico | 2750-1017F(MX) | 00095.102 | 9/7/00 | |
| Europe | 2750-1017F(EP) | 00095.101 | 9/7/00 | |
| United States | 2750-1017P | 00095.002 | 9/7/00 | |
| United States | 2750-1018P | 00096.002 | 9/8/00 | |
| Canada | 2750-1018F(CA) | 00096.100 | 9/8/00 | |
| Europe | 2750-1018F(EP) | 00096.101 | 9/8/00 | |
| Mexico | 2750-1018F(MX) | 00096.102 | 9/8/00 | |
| United States | 2750-1230P | 91090.001 | 9/13/00 | |
| United States | 2750-1019P | 00098.002 | 9/13/00 | |
| Canada | 2750-1019F(CA) | 00098.100 | 9/13/00 | |
| Europe | 2750-1019F(EP) | 00098.101 | 9/13/00 | |
| Mexico | 2750-1019F(MX) | 00098.102 | 9/13/00 | |
| United States | 2750-1229P | 80310.001 | 9/13/00 | |
| Europe | 2750-1021F(EP) | 00101.101 | 9/15/00 | |
| United States | 2750-1021P | 00101.002 | 9/15/00 | |
| Canada | 2750-1021F(CA) | 00101.100 | 9/15/00 | |
| Canada | 2750-1020F(CA) | 00099.100 | 9/15/00 | |
| Mexico | 2750-1021F(MX) | 00101.102 | 9/15/00 | |
| United States | 2750-1020P | 00099.002 | 9/15/00 | |
| Mexico | 2750-1020F(MX) | 00099.102 | 9/15/00 | |
| Europe | 2750-1020F(EP) | 00099.101 | 9/15/00 | |
| United States | 2750-1233P | 80312.001 | 9/18/00 | |
| United States | 2750-1234P | 91092.001 | 9/18/00 | |

All applications listed in the table above are expressly incorporated herein by reference in their entirety and for all purposes.

The SDFs of the invention can also be used as probes to search for genes that are related to the SDF within a species. Such related genes are typically considered to be members of a gene family. In such a case, the sequence similarity will often be concentrated into one or a few fragments of the sequence. The fragments of similar sequence that define

the gene family typically encode a fragment of a protein or RNA that has an enzymatic or structural function. The percentage of identity in the amino acid sequence of the domain that defines the gene family is preferably at least 70%, more preferably 80 to 95%, most preferably 85 to 99%. To search for members of a gene family within a species, a low stringency hybridization is usually performed, but this will depend upon the size, distribution and degree of sequence divergence of domains that define the gene family. SDFs encompassing regulatory regions can be used to identify coordinately expressed genes by using the regulatory region sequence of the SDF as a probe.

In the instances where the SDFs are identified as being expressed from genes that confer a particular phenotype, then the SDFs can also be used as probes to assay plants of different species for those phenotypes.

I.C. Methods to Inhibit Gene Expression

The nucleic acid molecules of the present invention can be used to inhibit gene transcription and/or translation. Example of such methods include, without limitation:

Antisense Constructs;

Ribozyme Constructs;

Chimeraplast Constructs;

Co-Suppression;

Transcriptional Silencing; and

Other Methods of Gene Expression.

C.1 Antisense

In some instances it is desirable to suppress expression of an endogenous or exogenous gene. A well-known instance is the FLAVOR-SAVOR™ tomato, in which the gene encoding ACC synthase is inactivated by an antisense approach, thus delaying softening of the fruit after ripening. See for example, U.S. Patent No. 5,859,330; U.S. Patent No. 5,723,766; Oeller, et al, *Science,* 254:437-439(1991); and Hamilton et al, *Nature*, 346:284-287 (1990). Also, timing of flowering can be controlled by suppression of the *FLOWERING LOCUS C* (*FLC*); high levels of this transcript are associated with late flowering, while absence of *FLC* is associated with early flowering (S.D. Michaels et al., *Plant Cell* 11:949 (1999). Also, the transition of apical meristem from production of leaves with associated shoots to flowering is regulated by *TERMINAL FLOWER1, APETALA1* and *LEAFY*. Thus,

when it is desired to induce a transition from shoot production to flowering, it is desirable to suppress *TFL1* expression (S.J. Liljegren, *Plant Cell* 11:1007 (1999)). As another instance, arrested ovule development and female sterility result from suppression of the ethylene forming enzyme but can be reversed by application of ethylene (D. De Martinis et al., *Plant Cell* 11:1061 (1999)). The ability to manipulate female fertility of plants is useful in increasing fruit production and creating hybrids.

In the case of polynucleotides used to inhibit expression of an endogenous gene, the introduced sequence need not be perfectly identical to a sequence of the target endogenous gene. The introduced polynucleotide sequence will typically be at least substantially identical to the target endogenous sequence.

Some polynucleotide SDFs in Table 1 represent sequences that are expressed in corn,wheat, rice, soybean *Arabidopsis* and/or other plants. Thus the invention includes using these sequences to generate antisense constructs to inhibit translation and/or degradation of transcripts of said SDFs, typically in a plant cell.

To accomplish this, a polynucleotide segment from the desired gene that can hybridize to the mRNA expressed from the desired gene (the "antisense segment") is operably linked to a promoter such that the antisense strand of RNA will be transcribed when the construct is present in a host cell. A regulated promoter can be used in the construct to control transcription of the antisense segment so that transcription occurs only under desired circumstances.

The antisense segment to be introduced generally will be substantially identical to at least a fragment of the endogenous gene or genes to be repressed. The sequence, however, need not be perfectly identical to inhibit expression. Further, the antisense product may hybridize to the untranslated region instead of or in addition to the coding sequence of the gene. The vectors of the present invention can be designed such that the inhibitory effect applies to other proteins within a family of genes exhibiting homology or substantial homology to the target gene.

For antisense suppression, the introduced antisense segment sequence also need not be full length relative to either the primary transcription product or the fully processed mRNA. Generally, a higher percentage of sequence identity can be used to compensate for the use of a shorter sequence. Furthermore, the introduced sequence need not have the same intron or exon pattern, and homology of non-coding segments may be equally effective. Normally, a sequence of between about 30 or 40 nucleotides and the full length of the transcript canbe used, though a sequence of at least about 100 nucleotides is preferred, a sequence of at least about 200 nucleotides is more preferred, and a sequence of at least about 500 nucleotides is especially preferred.

C.2. Ribozymes

It is also contemplated that gene constructs representing ribozymes and based on the SDFs in TABLE 1 are an object of the invention. Ribozymes can also be used to inhibit expression of genes by suppressing the translation of the mRNA into a polypeptide. It is possible to design ribozymes that specifically pair with virtually any target RNA and cleave the phosphodiester backbone at a specific location, thereby functionally inactivating the target RNA. In carrying out this cleavage, the ribozyme is not itself altered, and is thus capable of recycling and cleaving other molecules, making it a true enzyme. The inclusion of ribozyme sequences within antisense RNAs confers RNA-cleaving activity upon them, thereby increasing the activity of the constructs.

A number of classes of ribozymes have been identified. One class of ribozymes is derived from a number of small circular RNAs, which are capable of self-cleavage and replication in plants. The RNAs replicate either alone (viroid RNAs) or with a helper virus (satellite RNAs). Examples include RNAs from avocado sunblotch viroid and the satellite RNAs from tobacco ringspot virus, lucerne transient streak virus, velvet tobacco mottle virus, solanum nodiflorum mottle virus and subterranean clover mottle virus. The design and use of target RNA-specific ribozymes is described in Haseloff et al. *Nature,* 334:585 (1988).

Like the antisense constructs above, the ribozyme sequence fragment necessary for pairing need not be identical to the target nucleotides to be cleaved, nor identical to the sequences in TABLE 1. Ribozymes may be constructed by combining the ribozyme sequence and some fragment of the target gene which would allow recognition of the target gene mRNA by the resulting ribozyme molecule. Generally, the sequence in the ribozyme capable of binding to the target sequence exhibits a percentage of sequence identity with at least 80%, preferably with at least 85%, more preferably with at least 90% and most preferably with at least 95%, even more preferably, with at least 96%, 97%, 98% or 99% sequence identity to some fragment of a sequence in TABLE 1 or the complement thereof. The ribozyme can be equally effective in inhibiting mRNA translation by cleaving either in the untranslated or coding regions. Generally, a higher percentage of sequence identity can be used to compensate for the use of a shorter sequence. Furthermore, the introduced sequence need not have the same intron or exon pattern, and homology of non-coding segments may be equally effective.

C.3. Chimeraplasts

The SDFs of the invention, such as those described by Table 1, can also be used to construct chimeraplasts that can be introduced into a cell to produce at least one specific nucleotide change in a sequence corresponding to the SDF of the invention. A chimeraplast is an oligonucleotide comprising DNA and/or RNA that specifically hybridizes to a target region in a manner which creates a mismatched base-pair. This mismatched base-pair signals the cell's repair enzyme machinery which acts on the mismatched region resulting in the replacement, insertion or deletion of designated nucleotide(s). The altered sequence is then expressed by the cell's normal cellular mechanisms. Chimeraplasts can be designed to repair mutant genes, modify genes, introduce site-specific mutations, and/or act to interrupt or alter normal gene function (US Pat. Nos. 6,010,907 and 6,004,804; and PCT Pub. No. WO99/58723 and WO99/07865).

C.4. Sense Suppression

The SDFs of Table 1 of the present invention are also useful to modulate gene expression by sense suppression. Sense suppression represents another method of gene suppression by introducing at least one exogenous copy or fragment of the endogenous sequence to be suppressed.

Introduction of expression cassettes in which a nucleic acid is configured in the sense orientation with respect to the promoter into the chromosome of a plant or by a self-replicating virus has been shown to be an effective means by which to induce degradation of mRNAs of target genes. For an example of the use of this method to modulate expression of endogenous genes *see*, Napoli et al., *The Plant Cell* 2:279 (1990), and U.S. Patents Nos. 5,034,323, 5,231,020, and 5,283,184. Inhibition of expression may require some transcription of the introduced sequence.

For sense suppression, the introduced sequence generally will be substantially identical to the endogenous sequence intended to be inactivated. The minimal percentage of sequence identity will typically be greater than about 65%, but a higher percentage of sequence identity might exert a more effective reduction in the level of normal gene products. Sequence identity of more than about 80% is preferred, though about 95% to absolute identity would be most preferred. As with antisense regulation, the effect would likely apply to any other proteins within a similar family of genes exhibiting homology or substantial homology to the suppressing sequence.

C.5. Transcriptional Silencing

The nucleic acid sequences of the invention, including the SDFs of Table 1, and fragments thereof, contain sequences that can be inserted into the genome of an organism resulting in transcriptional silencing. Such regulatory sequences need not be operatively linked to coding sequences to modulate transcription of a gene. Specifically, a promoter sequence without any other element of a gene can be introduced into a genome to transcriptionally silence an endogenous gene (see, for example, Vaucheret, H et al. (1998) The Plant Journal 16: 651-659). As another example, triple helices can be formed using oligonucleotides based on sequences from TABLE 1, fragments thereof, and substantially similar sequence thereto. The oligonucleotide can be delivered to the host cell and can bind to the promoter in the genome to form a triple helix and prevent transcription. An oligonucleotide of interest is one that can bind to the promoter and block binding of a transcription factor to the promoter. In such a case, the oligonucleotide can be complementary to the sequences of the promoter that interact with transcription binding factors.

C.6. Other Methods to Inhibit Gene Expression

Yet another means of suppressing gene expression is to insert a polynucleotide into the gene of interest to disrupt transcription or translation of the gene.

Low frequency homologous recombination can be used to target a polynucleotide insert to a gene by flanking the polynucleotide insert with sequences that are substantially similar to the gene to be disrupted. Sequences from TABLE 1, fragments thereof, and substantially similar sequence thereto can be used for homologous recombination.

In addition, random insertion of polynucleotides into a host cell genome can also be used to disrupt the gene of interest. Azpiroz-Leehan et al., *Trends in Genetics* 13:152 (1997). In this method, screening for clones from a library containing random insertions is preferred to identifying those that have polynucleotides inserted into the gene of interest. Such screening can be performed using probes and/or primers described above based on sequences from TABLE 1, fragments thereof, and substantially similar sequence thereto. The screening can also be performed by selecting clones or $R_1$ plants having a desired phenotype.

I.D. Methods of Functional Analysis

The constructs described in the methods under I.C. above can be used to determine the function of the polypeptide encoded by the gene that is targeted by the constructs.

Down-regulating the transcription and translation of the targeted gene in the host cell or organisms, such as a plant, may produce phenotypic changes as compared to a wild-type cell or organism. In addition, *in vitro* assays can be used to determine if any biological activity, such as calcium flux, DNA transcription, nucleotide incorporation, etc., are being modulated by the down-regulation of the targeted gene.

Coordinated regulation of sets of genes, e.g., those contributing to a desired polygenic trait, is sometimes necessary to obtain a desired phenotype. SDFs of the invention representing transcription activation and DNA binding domains can be assembled into hybrid transcriptional activators. These hybrid transcriptional activators can be used with their corresponding DNA elements (i.e., those bound by the DNA-binding SDFs) to effect coordinated expression of desired genes (J.J. Schwarz et al., *Mol. Cell. Biol.* 12:266 (1992), A. Martinez et al., *Mol. Gen. Genet.* 261:546 (1999)).

The SDFs of the invention can also be used in the two-hybrid genetic systems to identify networks of protein-protein interactions (L. McAlister-Henn et al., *Methods* 19:330 (1999), J.C. Hu et al., *Methods* 20:80 (2000), M. Golovkin et al., *J. Biol. Chem.* 274:36428 (1999), K. Ichimura et al., *Biochem. Biophys. Res. Comm.* 253:532 (1998)). The SDFs of the invention can also be used in various expression display methods to identify important protein-DNA interactions (e.g. B. Luo et al., *J. Mol. Biol.* 266:479 (1997)).

## I.E. Promoters

The SDFs of the invention are also useful as structural or regulatory sequences in a construct for modulating the expression of the corresponding gene in a plant or other organism, *e.g.* a symbiotic bacterium. For example, promoter sequences associated to SDFs of Table 1 of the present invention can be useful in directing expression of coding sequences either as constitutive promoters or to direct expression in particular cell types, tissues, or organs or in response to environmental stimuli.

With respect to the SDFs of the present invention a promoter is likely to be a relatively small portion of a genomic DNA (gDNA) sequence located in the first 2000 nucleotides upstream from an initial exon identified in a gDNA sequence or initial "ATG" or methionine codon or translational start site in a corresponding cDNA sequence. Such promoters are more

likely to be found in the first 1000 nucleotides upstream of an initial ATG or methionine codon or translational start site of a cDNA sequence corresponding to a gDNA sequence. In particular, the promoter is usually located upstream of the transcription start site. The fragments of a particular gDNA sequence that function as elements of a promoter in a plant cell will preferably be found to hybridize to gDNA sequences presented and described in Table 1 at medium or high stringency, relevant to the length of the probe and its base composition.

Promoters are generally modular in nature. Promoters can consist of a basal promoter that functions as a site for assembly of a transcription complex comprising an RNA polymerase, for example RNA polymerase II. A typical transcription complex will include additional factors such as $TF_{II}B$, $TF_{II}D$, and $TF_{II}E$. Of these, $TF_{II}D$ appears to be the only one to bind DNA directly. The promoter might also contain one or more enhancers and/or suppressors that function as binding sites for additional transcription factors that have the function of modulating the level of transcription with respect to tissue specificity and of transcriptional responses to particular environmental or nutritional factors, and the like.

Short DNA sequences representing binding sites for proteins can be separated from each other by intervening sequences of varying length. For example, within a particular functional module, protein binding sites may be constituted by regions of 5 to 60, preferably 10 to 30, more preferably 10 to 20 nucleotides. Within such binding sites, there are typically 2 to 6 nucleotides that specifically contact amino acids of the nucleic acid binding protein. The protein binding sites are usually separated from each other by 10 to several hundred nucleotides, typically by 15 to 150 nucleotides, often by 20 to 50 nucleotides. DNA binding sites in promoter elements often display dyad symmetry in their sequence. Often elements binding several different proteins, and/or a plurality of sites that bind the same protein, will be combined in a region of 50 to 1,000 basepairs.

Elements that have transcription regulatory function can be isolated from their corresponding endogenous gene, or the desired sequence can be synthesized, and recombined in constructs to direct expression of a coding region of a gene in a desired tissue-specific, temporal-specific or other desired manner of inducibility or suppression. When hybridizations are performed to identify or isolate elements of a promoter by hybridization to the long sequences presented in TABLE 1, conditions are adjusted to account for the above-described nature of promoters. For example short probes, constituting the element sought, are preferably used under low temperature and/or high salt conditions. When long probes, which might include several promoter elements are used, low to medium stringency conditions are preferred when hybridizing to promoters across species.

If a nucleotide sequence of an SDF, or part of the SDF, functions as a promoter or fragment of a promoter, then nucleotide substitutions, insertions or deletions that do not substantially affect the binding of relevant DNA binding proteins would be considered equivalent to the exemplified nucleotide sequence. It is envisioned that there are instances where it is desirable to decrease the binding of relevant DNA binding proteins to silence or down-regulate a promoter, or conversely to increase the binding of relevant DNA binding proteins to enhance or up-regulate a promoter and vice versa. In such instances, polynucleotides representing changes to the nucleotide sequence of the DNA-protein contact region by insertion of additional nucleotides, changes to identity of relevant nucleotides, including use of chemically-modified bases, or deletion of one or more nucleotides are considered encompassed by the present invention. In addition, fragments of the promoter sequences described by Table 1 and variants thereof can be fused with other promoters or fragments to facilitate transcription and/or transcription in specific type of cells or under specific conditions.

Promoter function can be assayed by methods known in the art, preferably by measuring activity of a reporter gene operatively linked to the sequence being tested for promoter function. Examples of reporter genes include those encoding luciferase, green fluorescent protein, GUS, neo, cat and bar.

I.F. UTRs and Junctions

Polynucleotides comprising untranslated (UTR) sequences and intron/exon junctions are also within the scope of the invention. UTR sequences include introns and 5' or 3' untranslated regions ( 5' UTRs or 3' UTRs). Fragments of the sequences shown in TABLE 1 can comprise UTRs and intron/exon junctions.

These fragments of SDFs, especially UTRs, can have regulatory functions related to, for example, translation rate and mRNA stability. Thus, these fragments of SDFs can be isolated for use as elements of gene constructs for regulated production of polynucleotides encoding desired polypeptides.

Introns of genomic DNA segments might also have regulatory functions. Sometimes regulatory elements, especially transcription enhancer or suppressor elements, are found within introns. Also, elements related to stability of heteronuclear RNA and efficiency of splicing and of transport to the cytoplasm for translation can be found in intron elements. Thus, these segments can also find use as elements of expression vectors intended for use to transform plants.

Just as with promoters UTR sequences and intron/exon junctions can vary from those shown in TABLE 1. Such changes from those sequences preferably will not affect the regulatory activity of the UTRs or intron/exon junction sequences on expression, transcription, or translation unless selected to do so. However, in some instances, down- or up-regulation of such activity may be desired to modulate traits or phenotypic or *in vitro* activity.

I.G. Coding Sequences

Isolated polynucleotides of the invention can include coding sequences that encode polypeptides comprising an amino acid sequence encoded by sequences in TABLE 1 or an amino acid sequence presented in TABLE 1.

A nucleotide sequence encodes a polypeptide if a cell (or a cell free *in vitro* system) expressing that nucleotide sequence produces a polypeptide having the recited amino acid sequence when the nucleotide sequence is transcribed and the primary transcript is subsequently processed and translated by a host cell (or a cell free *in vitro* system) harboring the nucleic acid. Thus, an isolated nucleic acid that encodes a particular amino acid sequence can be a genomic sequence comprising exons and introns or a cDNA sequence that represents the product of splicing thereof. An isolated nucleic acid encoding an amino acid sequence also encompasses heteronuclear RNA, which contains sequences that are spliced out during expression, and mRNA, which lacks those sequences.

Coding sequences can be constructed using chemical synthesis techniques or by isolating coding sequences or by modifying such synthesized or isolated coding sequences as described above.

In addition to coding sequences encoding the polypeptide sequences of TABLE 1, which are native to corn, *Arabidopsis*, soybean, rice, wheat, and other plants the isolated polynucleotides can be polynucleotides that encode variants, fragments, and fusions of those native proteins. Such polypeptides are described below in part II.

In variant polynucleotides generally, the number of substitutions, deletions or insertions is preferably less than 20%, more preferably less than 15%; even more preferably less than 10%, 5%, 3% or 1% of the number of nucleotides comprising a particularly exemplified sequence. It is generally expected that non-degenerate nucleotide sequence changes that result in 1 to 10, more preferably 1 to 5 and most preferably 1 to 3 amino acid insertions, deletions or substitutions will not greatly affect the function of an encoded polypeptide. The most preferred embodiments are those wherein 1 to 20, preferably 1 to 10, most preferably 1 to 5 nucleotides

are added to, deleted from and/or substituted in the sequences specifically disclosed in TABLE 1.

Insertions or deletions in polynucleotides intended to be used for encoding a polypeptide preferably preserve the reading frame. This consideration is not so important in instances when the polynucleotide is intended to be used as a hybridization probe.

## II. Polypeptides and Proteins

IIA. Native polypeptides and proteins

Polypeptides within the scope of the invention include both native proteins as well as variants, fragments, and fusions thereof. Polypeptides of the invention are those encoded by any of the six reading frames of sequences shown in TABLE 1, preferably encoded by the three frames reading in the 5' to 3' direction of the sequences as shown.

Native polypeptides include the proteins encoded by the sequences shown in TABLE 1. Such native polypeptides include those encoded by allelic variants.

Polypeptide and protein variants will exhibit at least 75% sequence identity to those native polypeptides of TABLE 1. More preferably; the polypeptide variants will exhibit at least 85% sequence identity; even more preferably, at least 90% sequence identity; more preferably at least 95%, 96%, 97%, 98%, or 99% sequence identity. Fragments of polypeptide or fragments of polypeptides will exhibit similar percentages of sequence identity to the relevant fragments of the native polypeptide. Fusions will exhibit a similar percentage of sequence identity in that fragment of the fusion represented by the variant of the native peptide.

Furthermore, polypeptide variants will exhibit at least one of the functional properties of the native protein. Such properties include, without limitation, protein interaction, DNA interaction, biological activity, immunological activity, receptor binding, signal transduction, transcription activity, growth factor activity, secondary structure, three-dimensional structure, etc. As to properties related to *in vitro* or *in vivo* activities, the variants preferably exhibit at least 60% of the activity of the native protein; more preferably at least 70%, even more preferably at least 80%, 85%, 90% or 95% of at least one activity of the native protein.

One type of variant of native polypeptides comprises amino acid substitutions, deletions and/or insertions. Conservative substitutions are preferred to maintain the function or activity of the polypeptide.

Within the scope of percentage of sequence identity described above, a polypeptide of the invention may have additional individual amino acids or amino acid sequences inserted into

the polypeptide in the middle thereof and/or at the N-terminal and/or C-terminal ends thereof. Likewise, some of the amino acids or amino acid sequences may be deleted from the polypeptide.

A.1 Antibodies

Isolated polypeptides can be utilized to produce antibodies. Polypeptides of the invention can generally be used, for example, as antigens for raising antibodies by known techniques. The resulting antibodies are useful as reagents for determining the distribution of the antigen protein within the tissues of a plant or within a cell of a plant. The antibodies are also useful for examining the production level of proteins in various tissues, for example in a wild-type plant or following genetic manipulation of a plant, by methods such as Western blotting.

Antibodies of the present invention, both polyclonal and monoclonal, may be prepared by conventional methods. In general, the polypeptides of the invention are first used to immunize a suitable animal, such as a mouse, rat, rabbit, or goat. Rabbits and goats are preferred for the preparation of polyclonal sera due to the volume of serum obtainable, and the availability of labeled anti-rabbit and anti-goat antibodies as detection reagents. Immunization is generally performed by mixing or emulsifying the protein in saline, preferably in an adjuvant such as Freund's complete adjuvant, and injecting the mixture or emulsion parenterally (generally subcutaneously or intramuscularly). A dose of 50-200 μg/injection is typically sufficient. Immunization is generally boosted 2-6 weeks later with one or more injections of the protein in saline, preferably using Freund's incomplete adjuvant. One may alternatively generate antibodies by *in vitro* immunization using methods known in the art, which for the purposes of this invention is considered equivalent to *in vivo* immunization.

Polyclonal antisera is obtained by bleeding the immunized animal into a glass or plastic container, incubating the blood at 25°C for one hour, followed by incubating the blood at 4°C for 2-18 hours. The serum is recovered by centrifugation (e.g., 1,000xg for 10 minutes). About 20-50 ml per bleed may be obtained from rabbits.

Monoclonal antibodies are prepared using the method of Kohler and Milstein, *Nature* 256: 495 (1975), or modification thereof. Typically, a mouse or rat is immunized as described above. However, rather than bleeding the animal to extract serum, the spleen (and optionally several large lymph nodes) is removed and dissociated into single cells. If desired, the spleen cells can be screened (after removal of nonspecifically adherent cells) by applying a cell suspension to a plate, or well, coated with the protein antigen. B-cells producing membrane-

bound immunoglobulin specific for the antigen bind to the plate, and are not rinsed away with the rest of the suspension. Resulting B-cells, or all dissociated spleen cells, are then induced to fuse with myeloma cells to form hybridomas, and are cultured in a selective medium (e.g., hypoxanthine, aminopterin, thymidine medium, "HAT"). The resulting hybridomas are plated by limiting dilution, and are assayed for the production of antibodies which bind specifically to the immunizing antigen (and which do not bind to unrelated antigens). The selected Mab-secreting hybridomas are then cultured either *in vitro* (*e.g.*, in tissue culture bottles or hollow fiber reactors), or *in vivo* (as ascites in mice).

Other methods for sustaining antibody-producing B-cell clones, such as by EBV transformation, are known.

If desired, the antibodies (whether polyclonal or monoclonal) may be labeled using conventional techniques. Suitable labels include fluorophores, chromophores, radioactive atoms (particularly $^{32}P$ and $^{125}I$), electron-dense reagents, enzymes, and ligands having specific binding partners. Enzymes are typically detected by their activity. For example, horseradish peroxidase is usually detected by its ability to convert 3,3',5,5'-tetramethylbenzidine (TNB) to a blue pigment, quantifiable with a spectrophotometer.

A.2 In Vitro Applications of Polypeptides

Some polypeptides of the invention will have enzymatic activities that are useful *in vitro*. For example, the soybean trypsin inhibitor (Kunitz) family is one of the numerous families of proteinase inhibitors. It comprises plant proteins which have inhibitory activity against serine proteinases from the trypsin and subtilisin families, thiol proteinases and aspartic proteinases. Thus, these peptides find *in vitro* use in protein purification protocols and perhaps in therapeutic settings requiring topical application of protease inhibitors.

Delta-aminolevulinic acid dehydratase (EC 4.2.1.24) (ALAD) catalyzes the second step in the biosynthesis of heme, the condensation of two molecules of 5-aminolevulinate to form porphobilinogen and is also involved in chlorophyll biosynthesis(Kaczor et al. (1994) Plant Physiol. 1-4: 1411-7; Smith (1988) Biochem. J. 249: 423-8; Schneider (1976) Z. naturforsch. [C] 31: 55-63). Thus, ALAD proteins can be used as catalysts in synthesis of heme derivatives. Enzymes of biosynthetic pathways generally can be used as catalysts for *in vitro* synthesis of the compounds representing products of the pathway.

Polypeptides encoded by SDFs of the invention can be engineered to provide purification reagents to identify and purify additional polypeptides that bind to them. This

allows one to identify proteins that function as multimers or elucidate signal transduction or metabolic pathways. In the case of DNA binding proteins, the polypeptide can be used in a similar manner to identify the DNA determinants of specific binding (S. Pierrou et al., *Anal. Biochem.* 229:99 (1995), S. Chusacultanachai et al., *J. Biol. Chem.* 274:23591 (1999), Q. Lin et al., *J. Biol. Chem.* 272:27274 (1997)).

## II.B . POLYPEPTIDE VARIANTS , FRAGMENTS, AND FUSIONS

Generally, variants , fragments, or fusions of the polypeptides encoded by the SDFs of the invention can exhibit at least one of the activities of the identified domains and/or related polypeptides described in Table 1 corresponding to the SDF of interest.

### II.B .(1) Variants

A type of variant of the native polypeptides comprises amino acid substitutions. Conservative substitutions, described above (see II.), are preferred to maintain the function or activity of the polypeptide. Such substitutions include conservation of charge, polarity, hydrophobicity, size, etc. For example, one or more amino acid residues within the sequence can be substituted with another amino acid of similar polarity that acts as a functional equivalent, for example providing a hydrogen bond in an enzymatic catalysis. Substitutes for an amino acid within an exemplified sequence are preferably made among the members of the class to which the amino acid belongs. For example, the nonpolar (hydrophobic) amino acids include alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan and methionine. The polar neutral amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine, and glutamine. The positively charged (basic) amino acids include arginine, lysine and histidine. The negatively charged (acidic) amino acids include aspartic acid and glutamic acid.

Within the scope of percentage of sequence identity described above, a polypeptide of the invention may have additional individual amino acids or amino acid sequences inserted into the polypeptide in the middle thereof and/or at the N-terminal and/or C-terminal ends thereof. Likewise, some of the amino acids or amino acid sequences may be deleted from the polypeptide. Amino acid substitutions may also be made in the sequences; conservative substitutions being preferred.

One preferred class of variants are those that comprise (1) the domain of an encoded polypeptide and/or (2) residues conserved between the encoded polypeptide and related polypeptides. For this class of variants, the encoded polypeptide sequence is changed

by insertion, deletion, or substitution at positions flanking the domain and/or conserved residues.

Another class of variants includes those that comprise an encoded polypeptide sequence that is changed in the domain or conserved residues by a conservative substitution.

Yet another class of variants includes those that lack one of the *in vitro* activities, or structural features of the encoded polypeptides. One example is polypeptides or proteins produced from genes comprising dominant negative mutations. Such a variant may comprise an encoded polypeptide sequence with non-conservative changes in a particular domain or group of conserved residues.

II.A.(2) FRAGMENTS

Fragments of particular interest are those that comprise a domain identified for a polypeptide encoded by an SDF of the instant invention and variants thereof. Also, fragments that comprise at least one region of residues conserved between an SDF encoded polypeptide and its related polypeptides are of great interest. Fragments are sometimes useful as polypeptides corresponding to genes comprising dominant negative mutations are.

II.A.(3)FUSIONS

Of interest are chimeras comprising (1) a fragment of the SDF encoded polypeptide or variants thereof of interest and (2) a fragment of a polypeptide comprising the same domain. For example, an AP2 helix encoded by a SDF of the invention fused to second AP2 helix from ANT protein, which comprises two AP2 helices. The present invention also encompasses fusions of SDF encoded polypeptides, variants, or fragments thereof fused with related proteins or fragments thereof.

**DEFINITION OF DOMAINS**

The polypeptides of the invention may possess identifying domains. In addition, the domains within the SDF encoded polypeptide can be defined by the region that exhibits at least 70% sequence identity with the consensus sequences listed in the detailed description below of each of the domains.

The majority of the protein domain descriptions given below are obtained from Prosite, (http//www.expasy.ch/prosite/), and Pfam, (http//pfam.wustl.edu/browse.shtml).

1. (AAA) AAA-protein family signature

A large family of ATPases has been described [1 to 5] whose key feature is that they share a conserved region of about 220 amino acids that contains anATP-binding site. This family is now called AAA, for 'A'TPases 'A'ssociated with diverse cellular 'A'ctivities. The proteins that belong to this family either contain one or two AAA domains. Proteins containing two AAA domains:

- Mammalian and drosophila NSF (N-ethylmaleimide-sensitive fusion protein) and the fungal homolog, SEC18. These proteins are involved in intracellular transport between the endoplasmic reticulum and Golgi, as well as between different Golgi cisternae.
- Mammalian transitional endoplasmic reticulum ATPase (previously known as p97 or VCP) which is involved in the transfer of membranes from the endoplasmic reticulum to the golgi apparatus. This protein forms a ring-shaped homooligomer composed of six subunits. The yeast homolog is CDC48 and it may play a role in spindle pole proliferation.
- Yeast protein PAS1, essential for peroxisome assembly and the related protein PAS1 from Pichia pastoris.
- Yeast protein AFG2.
- Sulfolobus acidocaldarius protein SAV and Halobacterium salinarium cdcH which may be part of a transduction pathway connecting light to cell division.

Proteins containing a single AAA domain:

- Escherichia coli and other bacteria ftsH (or hflB) protein. FtsH is an ATP-dependent zinc metallopeptidase that seems to degrade the heat-shock sigma-32 factor.

It is an integral membrane protein with a large cytoplasmic C-terminal domain that contain both the AAA and the protease domains.

- Yeast protein YME1, a protein important for maintaining the integrity of the mitochondrial compartment. YME1 is also a zinc-dependent protease.
- Yeast protein AFG3 (or YTA10). This protein also seems to contain a AAA domain followed by a zinc-dependent protease domain.

Subunits from the regulatory complex of the 26S proteasome [6] which is involved in the ATP-dependent degradation of ubiquitinated proteins:

a) Mammalian subunit 4 and homologs in other higher eukaryotes, in yeast (gene YTA5) and fission yeast (gene mts2).

b) Mammalian subunit 6 (TBP7) and homologs in other higher eukaryotes and in yeast (gene YTA2).

c) Mammalian subunit 7 (MSS1) and homologs in other higher eukaryotes and in yeast (gene CIM5 or YTA3).

d) Mammalian subunit 8 (P45) and homologs in other higher eukaryotes and in yeast (SUG1 or CIM3 or TBY1) and fission yeast (gene let1).

Other probable subunits such as human TBP1 which seems to influences HIV gene expression by interacting with the virus tat transactivator protein and yeast YTA1 and YTA6.

- Yeast protein BCS1, a mitochondrial protein essential for the expression of the Rieske iron-sulfur protein.
- Yeast protein MSP1, a protein involved in intramitochondrial sorting of proteins.
- Yeast protein PAS8, and the corresponding proteins PAS5 from Pichia pastoris and PAY4 from Yarrowia lipolytica.
- Mouse protein SKD1 and its fission yeast homolog (SpAC2G11.06).
- Caenorhabditis elegans meiotic spindle formation protein mei-1.
- Yeast protein SAP1.
- Yeast protein YTA7.
- Mycobacterium leprae hypothetical protein A2126A.

It is proposed that, in general, the AAA domains in these proteins act as ATP-dependent protein clamps [5]. In addition to the ATP-binding 'A' and 'B' motifs, which are located in the N-terminal half of this domain, there is a highly conserved region located in the central part of the domain which was used to develop a signature pattern.

Consensus pattern: ~~[LIVMT]~~[LIVMT (SEQ ID NO: 518)]-x-~~[LIVMT]~~[LIVMT (SEQ ID NO: 518)]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x-~~[GATMC]~~[GATMC (SEQ ID NO: 178)]-[ST]-[NS]-x(4)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]- D-x-A-~~[LIFA]~~[LIFA (SEQ ID NO: 334)]-x-R

[1] Froehlich K.-U., Fries H.W., Ruediger M., Erdmann R., Botstein D., Mecke D. J. Cell Biol. 114:443-453(1991).

[2] Erdmann R., Wiebel F.F., Flessau A., Rytka J., Beyer A., Froehlich K.-U., Kunau W.-H. Cell 64:499-510(1991).

[3] Peters J.-M., Walsh M.J., Franke W.W. EMBO J. 9:1757-1767(1990).

[4] Kunau W.-H., Beyer A., Goette K., Marzioch M., Saidowsky J., Skaletz-Rorowski A., Wiebel F.F. Biochimie 75:209-224(1993).

[5] Confalonieri F., Duguet M. BioEssays 17:639-650(1995).[ 6] Hilt W., Wolf D.H. Trends Biochem. Sci. 21:96-102(1996).

2. ABC Membrane (ABC transporter transmembrane region). This family represents a unit of six transmembrane helices. Many members of the ABC transporter family (ABC_tran)have two such regions. See also descriptions of ABC Tran, below, and ABC2 membrane, above.

3. (ABC Tran) ABC transporters family signature. On the basis of sequence similarities a family of related ATP-bindingproteins has been characterized [1 to 5]. These proteins are associated with avariety of distinct biological processes in both prokaryotes and eukaryotes, but a majority of them are involved in active transport of small hydrophilic molecules across the cytoplasmic membrane. All these proteins share a conserved domain of some two hundred amino acid residues, which includes an ATP-binding site. These proteins are collectively known as ABC transporters. Proteins known to belong to this family are listed below (references are only provided for recently determined sequences).In prokaryotes: - Active transport systems components: alkylphosphonate uptake(phnC/phnK/ phnL); arabinose (araG); arginine (artP); dipeptide (dciAD;dppD/dppF); ferric enterobactin (fepC); ferrichrome (fhuC); galactoside (mglA); glutamine (glnQ); glycerol-3-phosphate (ugpC); glycine betaine/L-proline (proV); glutamate/aspatate (gltL); histidine (hisP); iron(III) (sfuC), iron(III) dicitrate (fecE); lactose (lacK); leucine/isoleucine/valine (braF/braG;livF/livG); maltose (malK); molybdenum (modC); nickel (nikD/ nikE); oligopeptide (amiE/amiF;oppD/oppF); peptide (sapD/sapF); phosphate (pstB); putrescine (potG); ribose (rbsA); spermidine/putrescine (potA); sulfate (cysA); vitamin B12 (btuD). - Hemolysin/leukotoxin export proteins hlyB, cyaB and lktB. - Colicin V export protein cvaB. - Lactococcin export protein lcnC [6]. - Lantibiotic transport proteins nisT (nisin) and spaT (subtilin). - Extracellular proteases B and C export protein prtD. - Alkaline protease secretion protein aprD. - Beta-(1,2)-glucan export proteins chvA and ndvA. - Haemophilus influenzae capsule-polysaccharide export protein bexA. - Cytochrome c biogenesis proteins ccmA (also known as cycV and helA). - Polysialic acid transport protein kpsT. - Cell division associated ftsE protein (function unknown). - Copper processing protein nosF from Pseudomonas stutzeri. - Nodulation protein nodI from Rhizobium (function unknown). - Escherichia coli proteins cydC and cydD. - Subunit A of the ABC excision nuclease (gene uvrA). - Erythromycin resistance protein from Staphylococcus epidermidis (gene msrA). - Tylosin

resistance protein from Streptomyces fradiae (gene tlrC) [7]. - Heterocyst differentiation protein (gene hetA) from Anabaena PCC 7120. - Protein P29 from Mycoplasma hyorhinis, a probable component of a high affinity transport system. - yhbG, a putative protein whose gene is linked with ntrA in many bacteria such as Escherichia coli, Klebsiella pneumoniae, Pseudomonas putida, Rhizobium meliloti and Thiobacillus ferrooxidans. - Escherichia coli and related bacteria hypothetical proteins yabJ, yadG, yagC, ybbA, ycjW, yddA, yehX, yejF, yheS, yhiG, yhiH, yjcW, yjjK, yojI, yrbF and ytfR.In eukaryotes: - The multidrug transporters (Mdr) (P-glycoprotein), a family of closely related proteins which extrude a wide variety of drugs out of the cell (for a review see [8]). - Cystic fibrosis transmembrane conductance regulator (CFTR), which is most probably involved in the transport of chloride ions. - Antigen peptide transporters 1 (TAP1, PSF1, RING4, HAM-1, mtp1) and 2 (TAP2, PSF2, RING11, HAM-2, mtp2), which are involved in the transport of antigens from the cytoplasm to a membrane-bound compartment for association with MHC class I molecules. - 70 Kd peroxisomal membrane protein (PMP70). - ALDP, a peroxisomal protein involved in X-linked adrenoleukodystrophy [9]. - Sulfonylurea receptor [10], a putative subunit of the B-cell ATP-sensitive potassium channel. - Drosophila proteins white (w) and brown (bw), which are involved in the import of ommatidium screening pigments. - Fungal elongation factor 3 (EF-3). - Yeast STE6 which is responsible for the export of the a-factor pheromone. - Yeast mitochondrial transporter ATM1. - Yeast MDL1 and MDL2. - Yeast SNQ2. - Yeast sporidesmin resistance protein (gene PDR5 or STS1 or YDR1). - Fission yeast heavy metal tolerance protein hmt1. This protein is probably involved in the transport of metal-bound phytochelatins. - Fission yeast brefeldin A resistance protein (gene bfr1 or hba2). - Fission yeast leptomycin B resistance protein (gene pmd1). - mbpX, a hypothetical chloroplast protein from Liverwort. - Prestalk-specific protein tagB from slime mold. This protein consists of two domains: a N-terminal subtilase catalytic domain and a C-terminal ABC transporter domain.As a signature pattern for this class of proteins, a conserved region which is located between the 'A' and the 'B' motifs of the ATP-binding site was used.

Consensus pattern: ~~[LIVMFYC]~~[LIVMFYC (SEQ ID NO: 439)]-[SA]-~~[SAPGLVFYKQH]~~[SAPGLVFYKQH (SEQ ID NO: 671)]-G-~~[DENQMW]~~[DENQMW (SEQ ID NO: 44)]-~~[KRQASPCLIMFW]~~[KRQASPCLIMFW (SEQ ID NO: 320)]-~~[KRNQSTAVM]~~[KRNQSTAVM (SEQ ID NO: 316)]-~~[KRACLVM]~~[KRACLVM (SEQ ID NO: 287)]-~~[LIVMFYPAN]~~[LIVMFYPAN (SEQ ID NO: 450)]-{PHY}-~~[LIVMFW]~~[LIVMFW (SEQ ID NO: 431)]- ~~[SAGCLIVP]~~[SAGCLIVP (SEQ ID NO: 657)]-

~~{FYWHP}~~{FYWHP (SEQ ID NO: 780)}-~~{KRHP}~~{KRHP (SEQ ID NO: 782)}-~~[LIVMFYWSTA]~~[LIVMFYWSTA (SEQ ID NO: 482)] The ATP-binding region is duplicated in araG, mdl, msrA, rbsA, tlrC, uvrA, yejF, Mdr's, CFTR, pmd1 and in EF-3. In some of those proteins, the above pattern only detect one of the two copies of the domain. The proteins belonging to this family also contain one or two copies of the ATP-binding motifs 'A' and 'B'.

[ 1] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J. Bioenerg. Biomembr. 22:571-592(1990).

[ 2] Higgins C.F., Gallagher M.P., Mimmack M.M., Pearce S.R. BioEssays 8:111-116(1988).

[ 3] Higgins C.F., Hiles I.D., Salmond G.P.C., Gill D.R., Downie J.A., Evans I.J., Holland I.B., Gray L., Buckels S.D., Bell A.W., Hermodson M.A. Nature 323:448-450(1986).

[ 4] Doolittle R.F., Johnson M.S., Husain I., van Houten B., Thomas D.C., Sancar A. Nature 323:451-453(1986).

[ 5] Blight M.A., Holland I.B. Mol. Microbiol. 4:873-880(1990).

[ 6] Stoddard G.W., Petzel J.P., van Belkum M.J., Kok J., McKay L.L. Appl. Environ. Microbiol. 58:1952-1961(1992).

[ 7] Rosteck P.R. Jr., Reynolds P.A., Hershberger C.L. Gene 102:27-32(1991).

[ 8] Gottesman M.M., Pastan I. J. Biol. Chem. 263:12163-12166(1988).

[ 9] Valle D., Gaertner J. Nature 361:682-683(1993).

[10] Aguilar-Bryan L., Nichols C.G., Wechsler S.W., Clement J.P. IV, Boyd A.E. III, Gonzalez G., Herrera-Sosa H., Nguy K., Bryan J., Nelson D.A. Science 268:423-426(1995).

4. (ACBP)

Acyl-CoA-binding protein signature

Acyl-CoA-binding protein (ACBP) is a small (10 Kd) protein that binds medium- and long-chain acyl-CoA esters with very high affinity and may function as an intracellular carrier of acyl-CoA esters [1]. ACBP is also known as diazepam binding inhibitor (DBI) or endozepine (EP) because of its ability to displace diazepam from the benzodiazepine (BZD) recognition site located on the GABA type A receptor. It is therefore possible that this protein also acts as a neuropeptide to modulate the action of the GABA receptor [2].ACBP is a highly conserved protein of about 90 residues that has been so far found in vertebrates, insects and yeast.

ACBP is also related to the N-terminal section of a probable transmembrane protein of unknown function whichhas been found in mammals. As a signature pattern, the region that corresponds to residues 19 to 37 in mammalian ACBP was selected.

Consensus pattern: P-[STA]-x-[DEN]-x-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x(2)-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-Y-~~[GSTA]~~[GSTA (SEQ ID NO: 217)]-x-[FY]-K-Q-[STA](2)-x-G-

[ 1] Rose T.M., Schultz E.R., Todaro G.J. Proc. Natl. Acad. Sci. U.S.A. 89:11287-11291(1992).

[ 2] Costa E., Guidotti A. Life Sci. 49:325-344(1991).

5. (AIRS)

AIR synthase related proteins

This family includes Hydrogen expression/formation protein HypE, AIR synthases, FGAM synthase and selenide, water dikinase.

6. (AMP-binding)

Putative AMP-binding domain signature

It has been shown [1 to 5] that a number of prokaryotic and eukaryotic enzymes which all probably act via an ATP-dependent covalent binding of AMP to their substrate, share a region of sequence similarity. These enzymes are: - Insects luciferase (luciferin 4-monooxygenase). Luciferase produces light by catalyzing the oxidation of luciferin in presence of ATP and molecular oxygen. - Alpha-aminoadipate reductase from yeast (gene LYS2). This enzyme catalyzes the activation of alpha-aminoadipate by ATP-dependent adenylation and the reduction of activated alpha-aminoadipate by NADPH. - Acetate--CoA ligase (acetyl-CoA synthetase), an enzyme that catalyzes the formation of acetyl-CoA from acetate and CoA. - Long-chain-fatty-acid--CoA ligase, an enzyme that activates long-chain fatty acids for both the synthesis of cellular lipids and their degradation via beta-oxidation. - 4-coumarate--CoA ligase (4CL), a plant enzyme that catalyzes the formation of 4-

coumarate-CoA from 4-coumarate and coenzyme A; the branchpoint reactions between general phenylpropanoid metabolism and pathways leading to various specific end products. - O-succinylbenzoic acid--CoA ligase (OSB-CoA synthetase) (gene menE) [6], a bacterial enzyme involved in the biosynthesis of menaquinone (vitamin K2). - 4-Chlorobenzoate--CoA ligase (EC 6.2.1.-) (4-CBA--CoA ligase) [7], a Pseudomonas enzyme involved in the degradation of 4-CBA. - Indoleacetate--lysine ligase (IAA-lysine synthetase) [8], an enzyme from Pseudomonas syringae that converts indoleacetate to IAA-lysine. - Bile acid-CoA ligase (gene baiB) from Eubacterium strain VPI 12708 [4]. This enzyme catalyzes the ATP-dependent formation of a variety of C-24 bile acid-CoA. - Crotonobetaine/carnitine-CoA ligase (EC 6.3.2.-) from Escherichia coli (gene caiC). - L-(alpha-aminoadipyl)-L-cysteinyl-D-valine synthetase (ACV synthetase) from various fungi (gene acvA or pcbAB). This enzyme catalyzes the first step in the biosynthesis of penicillin and cephalosporin, the formation of ACV from the constituent amino acids. The amino acids seem to be activated by adenylation. It is a protein of around 3700 amino acids that contains three related domains of about 1000 amino acids. - Gramicidin S synthetase I (gene grsA) from Bacillus brevis. This enzyme catalyzes the first step in the biosynthesis of the cyclic antibiotic gramicidin S, the ATP-dependent racemization of phenylalanine - Tyrocidine synthetase I (gene tycA) from Bacillus brevis. The reaction carried out by tycA is identical to that catalyzed by grsA - Gramicidin S synthetase II (gene grsB) from Bacillus brevis. This enzyme is a multifunctional protein that activates and polymerizes proline, valine, ornithine and leucine. GrsB consists of four related domains. - Enterobactin synthetase components E (gene entE) and F (gene entF) from Escherichia coli. These two enzymes are involved in the ATP-dependent activation of respectively 2,3-dihydroxybenzoate and serine during enterobactin (enterochelin) biosynthesis. - Cyclic peptide antibiotic surfactin synthase subunits 1, 2 and 3 from Bacillus subtilis. Subunits 1 and 2 contains three related domains while subunit 3 only contains a single domain. - HC-toxin synthetase (gene HTS1) from Cochliobolus carbonum. This enzyme activates the four amino acids (Pro, L-Ala, D-Ala and 2-amino-9,10-epoxi-8-oxodecanoic acid) that make up HC-toxin, a cyclic tetrapeptide. HTS1 consists of four related domains.There are also some proteins, whose exact function is not yet known, but whichare, very probably, also AMP-binding enzymes. These proteins are: - ORA (octapeptide-repeat antigen), a Plasmodium falciparum protein whose function is not known but which shows a high degree of similarity with the above proteins. - AngR, a Vibrio anguillarum protein. AngR is thought to be a transcriptional activator which modulates the anguibactin (an iron-binding siderophore) biosynthesis gene cluster operon. But it is believed [9], that angR is not

a DNA-binding protein, but rather an enzyme involved in the biosynthesis of anguibactin. This conclusion is based on three facts: the presence of the AMP-binding domain; the size of angR (1048 residues), which is far bigger than any bacterial transcriptional protein; and the presence of a probable S-acyl thioesterase immediately downstream of angR. - A hypothetical protein in mmsB 3'region in Pseudomonas aeruginosa. - Escherichia coli hypothetical protein ydiD. - Yeast hypothetical protein YBR041w. - Yeast hypothetical protein YBR222c. - Yeast hypothetical protein YER147c.All these proteins contain a highly conserved region very rich in glycine, serine, and threonine which is followed by a conserved lysine. A parallel can be drawn between this type of domain and the G-x(4)-G-K-[ST] ATP-/GTP-binding 'P-loop' domain or the protein kinases G-x-G-x(2)-[SG]-x(10,20)-KATP-binding domains.

Consensus pattern: ~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x(2)-[STG]-~~[STAG]~~[STAG (SEQ ID NO: 690)]-G-[ST]-~~[STEI]~~[STEI (SEQ ID NO: 742)]-[SG]-x-~~[PASLIVM]~~[PASLIVM (SEQ ID NO: 592)]- [KR] In a majority of cases the residue that follows the Lys at the end of the pattern is a Gly.

[ 1] Toh H. Protein Seq. Data Anal. 4:111-117(1991).

[ 2] Smith D.J., Earl A.J., Turner G. EMBO J. 9:2743-2750(1990).

[ 3] Schroeder J. Nucleic Acids Res. 17:460-460(1989).

[ 4] Mallonee D.H., Adams J.L., Hylemon P.B. J. Bacteriol. 174:2065-2071(1992).

[ 5] Turgay K., Krause M., Marahiel M.A. Mol. Microbiol. 6:529-546(1992).

[ 6] Driscoll J.R., Taber H.W. J. Bacteriol. 174:5063-5071(1992).

[ 7] Babbitt P.C., Kenyon G.L., Matin B.M., Charest H., Sylvestre M., Scholten J.D., Chang K.-H., Liang P.-H., Dunaway-Mariano D. Biochemistry 31:5594-5604(1992).

[ 8] Farrell D.H., Mikesell P., Actis L.A., Crosa J.H. Gene 86:45-51(1990).

7. AP2 domain

This 60 amino acid residue domain can bind to DNA [1]. This domain is plant specific. Members of this family are suggested to be related to pyridoxal phosphate-binding domains such as found in aminotran_2 [3]. AP2 domains are also described in Jofuku et al., co-pending U.S. Patent applications 08/700,152, 08/879,827, 08/912,272, 09/026,039.

[1] Ohme-takagi M, Shinshi H; Plant Cell 1995;7:173-182.

[2] Weigel D; Plant Cell 1995;7:388-389.

[3] Mushegian AR, Koonin EV; Genetics 1996;144:817-828.

8. ARID

The ARID domain is an AT-Rich Interaction domain sharing structural homology to DNA replication and repair nucleases and polymerases.

[1] Herrscher RF, Kaplan MH, Lelsz DL, Das C, Scheuermann R, Tucker PW; Genes Dev 1995;9:3067-3082.

[2] Yuan YC, Whitson RH, Liu Q, Itakura K, Chen Y; Nat Struct Biol 1998;5:959-964.

9. (ATP synt)

ATP synthase gamma subunit signature

ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a componentof the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), and a catalytic core, called coupling factor CF(1). The former acts as a proton channel; the latter is composed of five subunits, alpha, beta, gamma, delta and epsilon. Subunit gamma is believed to be important in regulating ATPase activity and the flow of protons through the CF(0) complex. The best conserved region of the gamma subunit [3] is its C-terminus which seems to be essential for assembly and catalysis. As a signature pattern to detect ATPase gamma subunits, a14 residue conserved segment where the last amino acid is found one to three residues from the C-terminal extremity was used.

Consensus pattern: [IV]-T-x-E-x(2)-[DE]-x(3)-G-A-x-~~[SAKR]~~[SAKR (SEQ ID NO: 666)]-

Note: Pea chloroplast gamma and two Bacillus species gamma subunits are not detected by this motif.

[ 1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

[ 2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[ 3] Miki J., Maeda M., Mukohata Y., Futai M. FEBS Lett. 232:221-226(1988).

10. (ATP Synt A)

Synthase a subunit signature

ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria,and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), which acts as a proton channel, and a catalytic core, termed coupling factor CF(1).The CF(0) a subunit, also called protein 6, is a key component of the proton channel; it may play a direct role in translocating protons across the membrane. It is a highly hydrophobic protein that has been predicted to contain 8 transmembrane regions [3].Sequence comparison of a subunits from all available sources reveals very few conserved regions. The best conserved region is located in what is predicted to be the fifth transmembrane domain. This region contains three perfectly conserved residues: an arginine, a leucine and an asparagine. Mutagenesis experiments of ATPase activity. This region was selected as a signature pattern.

Consensus pattern: ~~[STAGN]~~[STAGN (SEQ ID NO: 705)]-x-~~[STAG]~~[STAG (SEQ ID NO: 690)]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-R-L-x-~~[SAGV]~~[SAGV (SEQ ID NO: 664)]-N-~~[LIVMT]~~[LIVMT (SEQ ID NO: 518)] [R is important for proton translocation]

[ 1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

[ 2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[ 3] Lewis M.L., Chang J.A., Simoni R.D. J. Biol. Chem. 265:10541-10550(1990).

[ 4] Cain B.D., Simoni R.D. J. Biol. Chem. 264:3292-3300(1989).

11. ATP synthase B

Part of the CF(0) (base unit) of the ATP synthase. The base unit is thought to translocate protons through membrane (inner membrane in mitochondria, thylakoid membrane in plants,

cytoplasmic membrane in bacteria). The B subunits are thought to interact with the stalk of the CF(1) subunits.

12. (ATP synt C)

ATP synthase c subunit signature

ATP synthase (proton-translocating ATPase) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria,and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), which acts as a proton channel, and a catalytic core, termed coupling factor CF(1).The CF(0) c subunit (also called protein 9, proteolipid, or subunit III) [3,4]is a highly hydrophobic protein of about 8 Kd which has been implicated in the proton-conducting activity of ATPase. Structurally subunit c consist of two long terminal hydrophobic regions, which probably span the membrane, and a central hydrophilic region. N,N'-dicyclohexylcarbodiimide (DCCD) can bind covalently to subunit c and thereby abolish the ATPase activity. DCCD binds to a specific glutamate or aspartate residue which is located in the middle ofthe second hydrophobic region near the C-terminus of the protein. A signature pattern which includes the DCCD-binding residue was derived.

Consensus pattern: ~~[GSTA]~~[GSTA (SEQ ID NO: 217)]-R-[NQ]-P-x(10)-~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)](2)-x(3)-~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)]-x-[DE] [D or E binds DCCD]

[ 1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

[ 2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[ 3] Ivaschenko A.T., Karpenyuk T.A., Ponomarenko S.V. Biokhimiia 56:406-419(1991).

[ 4] Recipon H., Perasso R., Adoutte A., Quetier F. J. Mol. Evol. 34:292-303(1992).

13. (ATP synt DE)

ATP synthase, Delta/Epsilon chain

Part of the ATP synthase CF(1). These subunits are part of the head unit of the ATP synthase. The subunits are called delta and epsilon in human and metozoan species but in bacterial species the delta (D) subunit is theequivalent to the Oligomycin sensitive subunit (OSCP) in metozoans.

14. (ATP synt ab)

ATP synthase alpha and beta subunits signature

ATP synthase (proton-translocating ATPase) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria,and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), and a catalytic core, called coupling factor CF(1). The former acts as a proton channel; the latter is composed of five subunits, alpha, beta, gamma, delta and epsilon. The sequences of subunits alpha and beta are related and both contain a nucleotide-binding site for ATP and ADP. The beta chain has catalytic activity, while the alpha chain is a regulatory subunit. Vacuolar ATPases [3] (V-ATPases) are responsible for acidifying a variety of intracellular compartments in eukaryotic cells. Like F-ATPases, they are oligomeric complexes of a transmembrane and a catalytic sector. The sequenceof the largest subunit of the catalytic sector (70 Kd) is related to that ofF-ATPase beta subunit, while a 60 Kd subunit, from the same sector, is related to the F-ATPases alpha subunit [4].Archaebacterial membrane-associated ATPases are composed of three subunits.The alpha chain is related to F-ATPases beta chain and the beta chain is related to F-ATPases alpha chain [4].A protein highly similar to F-ATPase beta subunits is found [5] in some bacterial apparatus involved in a specialized protein export pathway that proceeds without signal peptide cleavage. This protein is known as fliI in Bacillus and Salmonella, Spa47 (mxiB) in Shigella flexneri, HrpB6 in Xanthomonas campestris and yscN in Yersinia virulence plasmids.To detect these ATPase subunits, a segment of ten amino-acid residues, containing two conserved serines, as a signature pattern was selected. The first serine seems to be important for catalysis - in the ATPase alpha chain at least - as its mutagenesis causes catalytic impairment.

Consensus pattern: P-[SAP]-[LIV]-[DNH]-x(3)-S-x-S [The first S is a putative active site residue]

[ 1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

[ 2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[ 3] Nelson N. J. Bioenerg. Biomembr. 21:553-571(1989).

[ 4] Gogarten J.P., Kibak H., Dittrich P., Taiz L., Bowman E.J., Bowman B.J., Manolson M.F., Poole R.J., Date T., Oshima T., Konishi J., Denda K., Yoshida M. Proc. Natl. Acad. Sci. U.S.A. 86:6661-665(1989).

[ 5] Dreyfus G., Williams A.W., Kawagishi I., MacNab R.M. J. Bacteriol. 175:3131-3138(1993).

15. (ATP synt ab C)

ATP synthase ab C terminal.

Number of members: 190

[1] Abrahams JP, Leslie AG, Lutter R, Walker JE; "Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria." Nature 1994;370:621-628.

16. (A deaminase)

Adenosine and AMP deaminase signature

Adenosine deaminase catalyzes the hydrolytic deamination ofadenosine into inosine. AMP deaminase catalyzes the hydrolytic deamination of AMP into IMP. It has been shown [1] that these two types of enzymes share three regions of sequence similarities; these regions are centered on residues which are proposed to play an important role in the catalytic mechanism of these two enzymes. One of these regions, containing two conserved aspartic acid residues that are potential active site residues was selected.

Consensus pattern: [SA]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[NGS]-[STA]-D-D-P [The two D's are putative active site residues]

[ 1] Chang Z., Nygaard P., Chinault A.C., Kellems R.E. Biochemistry 30:2273-2280(1991).

17. (Acetyltransf)
Acetyltransferase (GNAT) family.

This family contains proteins with N-acetyltransferase functions.

[1] Neuwald AF, Landsman D; Trends Biochem Sci 1997;22:154-155.

18. (Aconitase C)
Aconitase family signature

Aconitase (aconitate hydratase) (EC 4.2.1.3) [1] is the enzyme from the tricarboxylic acid cycle that catalyzes the reversible isomerization of citrate and isocitrate. Cis-aconitate is formed as an intermediary product during the course of the reaction. In eukaryotes two isozymes of aconitase are known to exist: one found in the mitochondrial matrix and the other found in the cytoplasm. Aconitase, in its active form, contains a 4Fe-4S iron-sulfur cluster; three cysteine residues have been shown to be ligands of the 4Fe-4S cluster.It has been shown that the aconitase family also contains the followingproteins: - Iron-responsive element binding protein (IRE-BP). IRE-BP is a cytosolic protein that binds to iron-responsive elements (IREs). IREs are stem-loop structures found in the 5'UTR of ferritin, and delta aminolevulinic acid synthase mRNAs, and in the 3'UTR of transferrin receptor mRNA. IRE-BP also express aconitase activity. - 3-isopropylmalate dehydratase (EC 4.2.1.33) (isopropylmalate isomerase), the enzyme that catalyzes the second step in the biosynthesis of leucine. - Homoaconitase (EC 4.2.1.36) (homoaconitate hydratase), an enzyme that participates in the alpha-aminoadipate pathway of lysine biosynthesis and that converts cis-homoaconitate into homoisocitric acid. - Esherichia coli protein ybhJ.As a signature for proteins from the aconitase family, two conserved regions that contain the three cysteine ligands of the 4Fe-4Scluster were selected.

Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(2)-~~[GSACIVM]~~[GSACIVM (SEQ ID NO: 190)]-x-[LIV]-~~[GTIV]~~[GTIV (SEQ ID NO: 255)]-[STP]-C-x(0,1)-T-N-~~[GSTANI]~~[GSTANI (SEQ ID NO: 235)]-x(4)-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)] [C binds the iron-sulfur center]

Consensus pattern: G-x(2)-~~[LIVWPQ]~~[LIVWPQ (SEQ ID NO: 539)]-x(3)-[GAC]-C-~~[GSTAM]~~[GSTAM (SEQ ID NO: 232)]-~~[LIMPTA]~~[LIMPTA (SEQ ID NO: 343)]-C-~~[LIMV]~~[LIMV (SEQ ID NO: 347)]- [GA] [The two C's bind the iron-sulfur center]

[ 1] Gruer M.J., Artymiuk P.J., Guest J.R. Trends Biochem. Sci. 22:3-6(1997).

19. (Acyl-CoA dh)

Acyl-CoA dehydrogenases signatures

Acyl-CoA dehydrogenases [1,2,3] are enzymes that catalyze the alpha, beta-dehydrogenation of acyl-CoA esters and transfer electrons to ETF, the electron transfer protein. Acyl-CoA dehydrogenases are FAD flavoproteins. This family currently includes: - Five eukaryotic isozymes that catalyze the first step of the beta-oxidation cycles for fatty acids with various chain lengths. These are short (SCAD) (EC 1.3.99.2), medium (MCAD) (EC 1.3.99.3), long (LCAD) (EC 1.3.99.13), very-long (VLCAD) and short/branched (SBCAD) chain acyl-CoA dehydrogenases. These enzymes are located in the mitochondrion. They are all homotetrameric proteins of about 400 amino acid residues except VLCAD which is a dimer and which contains, in its mature form, about 600 residues. - Glutaryl-CoA dehydrogenase (EC 1.3.99.7) (GCDH), which is involved in the catabolism of lysine, hydroxylysine and tryptophan. - Isovaleryl-CoA dehydrogenase (EC 1.3.99.10) (IVD), involved in the catabolism of leucine. - Acyl-coA dehydrogenases acsA and mmgC from Bacillus subtilis. - Butyryl-CoA dehydrogenase (EC 1.3.99.2) from Clostridium acetobutylicum. - Escherichia coli protein caiA [4]. - Escherichia coli protein aidB. Two conserved regions were selected as signature patterns. The first is located in the center of these enzymes, the second in the C-terminal section.

Consensus pattern: [GAC]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[ST]-E-x(2)-~~[GSAN]~~[GSAN (SEQ ID NO: 205)]-G-[ST]-D-x(2)-[GSA]

Consensus pattern: [QDE]-x(2)-G-[GS]-x-G-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x(2)-[DEN]-x(4)-[KR]-x(3)- [DEN]

[ 1] Tanaka K., Ikeda, Matsubara Y., Hyman D.B. Enzyme 38:91-107(1987).

[ 2] Matsubara Y., Indo Y., Naito E., Ozasa H., Glassberg R., Vockley J., Ikeda Y., Kraus J., Tanaka K. J. Biol. Chem. 264:16321-16331(1989).

[ 3] Aoyama T., Ueno I., Kamijo T., Hashimoto T. J. Biol. Chem. 269:19088-19094(1994).

[ 4] Eichler K., Bourgis F., Buchet A., Kleber H.-P., Mandrand-Berthelot M.-A. Mol. Microbiol. 13:775-786(1994).

20. (Acyl transf)

Acyl transferase domain

Number of members: 161

[1] Serre L, Verbree EC, Dauter Z, Stuitje AR, Derewenda ZS; Medline: 95286570 "The Escherichia coli malonyl-CoA:acyl carrier protein transacylase at 1.5-A resolution. Crystal structure of a fatty acid synthase component." J Biol Chem 1995;270:12961-12964.

21. Acylphosphatase signatures

Acylphosphatase (EC 3.6.1.7) [1,2] catalyzes the hydrolysis of various acylphosphate carboxyl-phosphate bonds such as carbamyl phosphate, succinylphosphate, 1,3-diphosphoglycerate, etc. The physiological role of this enzymeis not yet clear. Acylphosphatase is a small protein of around 100 amino-acid residues. There are two known isozymes. One seems to be specific to muscular tissues, the other, called 'organ-common type', is found in many different tissues.While acylphosphatase have been so far only characterized in vertebrates,there are a number of bacterial and archebacterial hypothetical proteins that are highly similar to that enzyme and that probably possess the same activity.These proteins are: - Escherichia coli hypothetical protein yccX. - Bacillus subtilis hypothetical protein yflL. - Archaeoglobus fulgidus hypothetical protein AF0818. Two conserved regions were selected as signature patterns. The first is located in the N-terminal section, while the second is found in the central part ofthe protein sequence.

Consensus pattern: [LIV]-x-G-x-V-Q-G-V-x-[FM]-R

Consensus pattern: G-[FYW]-[AVC]-~~[KRQAM]~~[KRQAM (SEQ ID NO: 319)]-N-x(3)-G-x-V-x(5)-G

[ 1] Stefani M., Ramponi G. Life Chem. Rep. 12:271-301(1995).
[ 2] Stefani M., Taddei N., Ramponi G. Cell. Mol. Life Sci. 53:141-151(1997).

22. (Adap comp sub)

Clathrin adaptor complexes medium chain signatures.

Clathrin coated vesicles (CCV) mediate intracellular membrane traffic such asreceptor mediated endocytosis. In addition to clathrin, the CCV are composed of a number of other components including oligomeric complexes which are knownas adaptor or clathrin assembly proteins (AP) complexes [1]. The adaptor complexes are believed to interact with the cytoplasmic tails of membrane proteins, leading to their selection and concentration. In mammals two type of adaptor complexes are known: AP-1 which is associated with the Golgi complex and AP-2 which is associated with the plasma membrane. Both AP-1 and AP-2 are heterotetramers that consist of two large chains - the adaptins - (gamma and beta' in AP-1; alpha and beta in AP-2); a medium chain (AP47 in AP-1; AP50 inAP-2) and a small chain (AP19 in AP-1; AP17 in AP-2). The medium chains of AP-1 and AP-2 are evolutionary related proteins of about 50 Kd. Homologs of AP47 and AP50 have also been found in Caenorhabditis elegans (genes unc-101 and ap50) [2] and yeast (gene APM1 or YAP54) [3].Some more divergent, but clearly evolutionary related proteins have also been found in yeast: APM2 and YBR288c., Two conserved regions were selected as signature patterns, one located in the N-terminal region, the other from the central section of these proteins.

Consensus pattern: [IVT]-[GSP]-W-R-x(2,3)-[GAD]-x(2)-[HY]-x(2)-N-x-~~[LIVMAFY]~~[LIVMAFY (SEQ ID NO: 385)](3)-D-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-~~[LIVMT]~~[LIVMT (SEQ ID NO: 518)]-E

Consensus pattern: [LIV]-x-F-I-P-P-x-G-x-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x-L-x(2)-Y

[ 1] Pearse B.M., Robinson M.S. Annu. Rev. Cell Biol. 6:151-171(1990).

[ 2] Lee J., Jongeward G.D., Sternberg P.W. Genes Dev. 8:60-73(1994).

[ 3] Nakayama Y., Goebl M., O'Brine G.B., Lemmon S., Pingchang C.E., Kirchhausen T. Eur. J. Biochem. 202:569-574(1991).

23. (Adenylsucc synt)

Adenylosuccinate synthetase signatures

Adenylosuccinate synthetase (EC 6.3.4.4) [1] plays an important role in purinebiosynthesis, by catalyzing the GTP-dependent conversion of IMP and aspartic acid to AMP. Adenylosuccinate synthetase has been characterized from various sources ranging from Escherichia coli (gene purA) to vertebrate tissues. Invertebrates, two isozymes are present - one involved in purine biosynthesis and the other in the purine nucleotide cycle. Two conserved regions were selected as signature patterns. The first one is a perfectly conserved octapeptide located in the N-terminal section and which is involved in GTP-binding [2]. The second one includes a lysine residue known [2] to be essential for the enzyme's activity.

Consensus pattern: Q-W-G-D-E-G-K-G

Consensus pattern: G-I-[GR]-P-x-Y-x(2)-K-x(2)-R [K is the active site residue]

[ 1] Wiesmueller L., Wittbrodt J., Noegel A.A., Schleicher M. J. Biol. Chem. 266:2480-2485(1991).

[ 2] Silva M.M., Poland B.W., Hoffman C.R., Fromm H.J., Honzatko R.B. J. Mol. Biol. 254:431-446(1995).

[ 3] Bouyoub A., Barbier G., Forterre P., Labedan B. 2.3.CO;2-"J. Mol. Biol. 261:144-154(1996).

24. (AdoHcyase)

S-adenosyl-L-homocysteine hydrolase signatures

S-adenosyl-L-homocysteine hydrolase (EC 3.3.1.1) (AdoHcyase) is an enzyme of the activated methyl cycle, responsible for the reversible hydratation of S-adenosyl-L-homocysteine into adenosine and homocysteine. AdoHcyase is anubiquitous enzyme which binds and requires NAD+ as a cofactor. AdoHcyase is a highly conserved protein [1] of about 430 to 470 amino acids. Two highly conserved regions were selected as signature patterns. The first pattern is located in the N-terminal section; the second is derived from aglycine-rich region in the central part of AdoHcyase; a region thought to be involved in NAD-binding.

Consensus pattern: [GSA]-[CS]-N-x-~~[FYLM]~~[FYLM (SEQ ID NO: 130)]-S-[ST]-[QA]-[DEN]-x-[AV]-[AT]-[AD]- [AC]-~~[LIVMCG]~~[LIVMCG (SEQ ID NO: 398)]

Consensus pattern: [GA]-[KS]-x(3)-[LIV]-x-G-[FY]-G-x-[VC]-G-[KRL]-G-x-[ASC]

[ 1] Sganga M.W., Aksamit R.R., Cantoni G.L., Bauer C.E. Proc. Natl. Acad. Sci. U.S.A. 89:6328-6332(1992).

25. AhpC/TSA family

This family contains proteins related to alkyl hydroperoxide reductaseComment: (AhpC) and thiol specific antioxidant (TSA).

[1] Chae HZ, Robison K, Poole LB, Church G, Storz G, Rhee SG, Proc Natl Acad Sci U S A 1994;91:7017-7021

26. (Aldose epim)

Aldose 1-epimerase putative active site Aldose 1-epimerase (EC 5.1.3.3) (mutarotase) is the enzyme responsible for the anomeric interconversion of D-glucose and other aldoses between their alpha- and beta-forms. The sequence of mutarotase from two bacteria, Acinetobacter calcoaceticus and Streptococcus thermophilus is available [1]. It has also been shown that, on the basis of extensive sequence similarities, a mutarotase domain seem to be present in the C-terminal half of the fungal GAL10 protein which encodes, in the N-terminal part, for UDP-glucose 4-epimerase. The best conserved region in the sequence of

mutarotase is centered around a conserved histidine residue which may be involved in the catalytic mechanism.

Consensus pattern: [NS]-x-T-N-H-x-Y-[FW]-N-[LI]

[ 1] Poolman B., Royer T.J., Mainzer S.E., Schmidt B.F. J. Bacteriol. 172:4037-4047(1990).

27. (AlkA DNA repair)

Alkylbase DNA glycosidases alkA family signature

Alkylbase DNA glycosidases [1] are DNA repair enzymes that hydrolyzes the deoxyribose N-glycosidic bond to excise various alkylated bases from a damaged DNA polymer. In Escherichia coli there are two alkylbase DNA glycosidases: one (gene tag)which is constitutively expressed and which is specific for the removal of 3-methyladenine (EC 3.2.2.20), and one (gene alkA) which is induced during adaptation to alkylation and which can remove a variety of alkylation products (EC 3.2.2.21). Tag and alkA do not share any region of sequence similarity. In yeast there is an alkylbase DNA glycosidase (gene MAG1) [2,3], which can remove 3-methyladenine or 7-methyladenine and which is structurally related to alkA. MAG and alkA are both proteins of about 300 amino acid residues. While the C- and N-terminal ends appear to be unrelated, there is a central region of about 130 residues which is well conserved. A portion of this region has been selected as a signature pattern .

Consensus pattern: G-I-G-x-W-[ST]-[AV]-x-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)](2)-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(8)-[MF]-x(2)- [ED]-D

[ 1] Lindahl T., Sedgwick B. Annu. Rev. Biochem. 57:133-157(1988).

[ 2] Berdal K.G., Bjoras M., Bjelland S., Seeberg E.C. EMBO J. 9:4563-4568(1990).

[ 3] Chen J., Derfler B., Samson L. EMBO J. 9:4569-4575(1990).

28. Ammonium transporters signature

A number of proteins involved in the transport of ammonium ions across amembrane as well as some yet uncharacterized proteins have been shown [1,2] to be evolutionary related. These proteins are: - Yeast ammonium transporters MEP1, MEP2 and MEP3. - Arabidopsis thaliana high affinity ammonium transporter (gene AMT1). - Corynebacterium glutamicum ammonium and methylammonium transport system. - Escherichia coli putative ammonium transporter amtB. - Bacillus subtilis nrgA. - Mycobacterium tuberculosis hypothetical protein MtCY338.09c. - Synechocystis strain PCC 6803 hypothetical proteins sll0108, sll0537 and sll1017. - Methanococcus jannaschii hypothetical proteins MJ0058 and MJ1343. - Caenorhabditis elegans hypothetical proteins C05E11.4, F49E11.3 and M195.3. As expected by their transport function, these proteins are highly hydrophobic and seem to contain from 10 to 12 transmembrane domains. The best conserved region seems to be located in the fifth (or sixth) transmembrane region and is used as a signature pattern.

Consensus pattern: D-~~[FYWS]~~[FYWS (SEQ ID NO: 160)]-A-G-[GSC]-x(2)-[IV]-x(3)-[SAG](2)-x(2)-[SAG]- ~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x(3)-~~[LIVMFYWA]~~[LIVMFYWA (SEQ ID NO: 464)](2)-x-[GK]-x-R

[ 1] Ninnemann O., Janniaux J.-C., Frommer W.B. EMBO J. 13:3464-3471(1994).
[ 2] Siewe R.M., Weil B., Burkovski A., Eikmanns B.J., Eikmanns M., Kraemer R. J. Biol. Chem. 271:5398-5403(1996).
[ 3] Saier M.H. Jr. Adv. Microbiol. Physiol. 40:81-136(1998).

29. (Arch_histone)

CBF/NF-Y subunits signatures

Diverse DNA binding proteins are known to bind the CCAAT box, a common cis-acting element found in the promoter and enhancer regions of a large number of genes in eukaryotes. Amongst these proteins is one known as the CCAAT-binding factor (CBF) or NF-Y [1]. CBF is a heteromeric transcription factor that consists of two different components both needed for DNA-binding. The HAP protein complex of yeast binds to the upstream activation site of cytochrome C iso-1 gene (CYC1) as well as other genes involved in mitochondrial electron transport and activates their expression. It also recognizes the sequence CCAAT and is structurally and evolutionary related to CBF. The first subunit of

CBF, known as CBF-A or NF-YB in vertebrates, HAP3 in budding yeast and as php3 in fission yeast, is a protein of 116 to 210 amino-acid residues which contains a highly conserved central domain of about 90residues. This domain seems to be involved in DNA-binding; a signature pattern had been developed from its central part. The second subunit of CBF, known as CBF-B or NF-YA in vertebrates, HAP2 in budding yeast and php2 in fission yeast, is a protein of 265 to 350 amino-acid residues which contains a highly conserved region of about 60 residues. This region, called the 'essential core' [2], seems to consist of two subdomains: an N-terminal subunit-association domain and a C-terminal DNA recognition domain. A signature pattern has been developed from a section of the subunit-association domain.

Consensus pattern: C-V-S-E-x-I-S-F-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-T-[SG]-E-A-[SC]-[DE]-[KRQ]-C-

Consensus pattern: Y-V-N-A-K-Q-Y-x-R-I-L-K-R-R-x-A-R-A-K-L-E-

[ 1] Li X.-Y., Mantovani R., Hooft van Huijsduijnen R., Andre I., Benoist C., Mathis D. Nucleic Acids Res. 20:1087-1091(1992).

[ 2] Olesen J.T., Fikes J.D., Guarente L. Mol. Cell. Biol. 11:611-619(1991).

30. Argininosuccinate synthase signatures

Argininosuccinate synthase (EC 6.3.4.5) (AS) is a urea cycle enzyme that catalyzes the penultimate step in arginine biosynthesis: the ATP-dependent ligation of citrulline to aspartate to form argininosuccinate, AMP andpyrophosphate [1,2].In humans, a defect in the AS gene causes citrullinemia, a genetic disease characterized by severe vomiting spells and mental retardation.AS is a homotetrameric enzyme of chains of about 400 amino-acid residues. Anarginine seems to be important for the enzyme's catalytic mechanism. The sequences of AS from various prokaryotes, archaebacteria and eukaryotes show significant similarity. Two signature patterns have been selected for AS. The first is a highly conserved stretch of nine residues located in the N-terminal extremity of these enzymes, the second is derived from a conserved region which contains one of the conserved arginine residues.

Consensus pattern: [AS]-[FY]-S-G-G-[LV]-D-T-[ST]-

Consensus pattern: G-x-T-x-K-G-N-D-x(2)-R-F-

[ 1] van Vliet F., Crabeel M., Boyen A., Tricot C., Stalon V., Falmagne P., Nakamura Y., Baumberg S., Glansdorff N. Gene 95:99-104(1990).
[ 2] Morris C.J., Reeve J.N. J. Bacteriol. 170:3125-3130(1988).

31. Armadillo/beta-catenin-like repeats

Approx. 40 amino acid repeat. Tandem repeats form super-helix of helices that is proposed to mediate interaction of beta-catenin with its ligands. CAUTION: This family does not contain all known armadillo repeats.

[1] Huber AH, Nelson WJ, Weis WI, Cell 1997;90:871-882.
[2] Gumbiner BM, Curr Opin Cell Biol 1995;7:634-640.
[3] Cavallo R, Rubenstein D, Peifer M, Curr Opin Genet Dev 1997;7:459-466.
[4] Su LK, Vogelstein B, Kinzler KW, Science 1993;262:1734-1737.
[5] Masiarz FR, Munemitsu S, Polakis P Science 1993;262:1731-1734
[6] Peifer M, Wieschaus E, Cell 1990;63:1167-1176.

32. (Asn Synthase)

Asparagine synthase

This family is always found associated with GATase_2. Members of this family catalyse the conversion of aspartate to asparagine.

33. Asparaginase_2

Asparaginase 12 members

34. (Aspartyl tRNA N)

Aminoacyl-transfer RNA synthetases class-II signatures

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern: ~~[GSTALVF]~~[GSTALVF (SEQ ID NO: 231)]-~~{DENQHRKP}~~{DENQHRKP (SEQ ID NO: 775)}-~~[GSTA]~~[GSTA (SEQ ID NO: 217)]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-[DE]-R-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x-~~[LIVMSTAG]~~[LIVMSTAG (SEQ ID NO: 514)]-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]

[ 1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).

[ 2] Delarue M., Moras D. BioEssays 15:675-687(1993).

[ 3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).

[ 4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

[ 5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).

[ 6] Cusack S. Biochimie 75:1077-1081(1993).

[ 7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).

[ 8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

35. (ArfGap) Putative GTP-ase activating protein for Arf. Putative zinc fingers with GTPase activating proteins (GAPs) towards the small GTPase, Arf. The GAP of ARD1 stimulates GTPase hydrolysis for ARD1 but not ARFs. Number of members: 34

[1]Medline: 96324970. Identification and cloning of centaurin-alpha. A novel phosphatidylinositol 3,4,5-trisphosphate-binding protein from rat brain. Hammonds-Odie LP, Jackson TR, Profit AA, Blader IJ, Turck CW, Prestwich GD, Theibert AB; J Biol Chem 1996;271:18859-18868.

[2]Medline: 97296423. A target of phosphatidylinositol 3,4,5-trisphosphate with a zinc finger motif similar to that of the ADP-ribosylation -factor GTPase-activating protein and two pleckstrin homology domains. Tanaka K, Imajoh-Ohmi S, Sawada T, Shirai R, Hashimoto Y, Iwasaki S, Kaibuchi K, Kanaho Y, Shirai T, Terada Y, Kimura K, Nagata S, Fukui Y; Eur J Biochem 1997;245:512-519.

[3] 98112795. Molecular characterization of the GTPase-activating domain of ADP-ribosylation factor domain protein 1 (ARD1). Vitale N, Moss J, Vaughan M; J Biol Chem 1998;273:2553-2560.

36. Apolipoprotein. Apolipoprotein A1/A4/E family. This family includes: Swiss:P02647 Apolipoprotein A-I. Swiss:P06727 Apolipoprotein A-IV. Swiss:P02649 Apolipoprotein E. These proteins contain several 22 residue repeats which form a pair of alpha helices. Number of members: 42

[1]Medline: 91289138. Three-dimensional structure of the LDL receptor-binding domain of human apolipoprotein E. Wilson C, Wardell MR, Weisgraber KH, Mahley RW, Agard DA; Science 1991;252:1817-1822.

37. Amino acid permeases signature

Amino acid permeases are integral membrane proteins involved in the transport of amino acids into the cell. A number of such proteins have been found to be evolutionary related [1,2,3]. These proteins are: - Yeast general amino acid permeases (genes GAP1, AGP2 and AGP3). - Yeast basic amino acid permease (gene ALP1). - Yeast Leu/Val/Ile permease (gene BAP2). - Yeast arginine permease (gene CAN1). - Yeast dicarboxylic amino acid permease

(gene DIP5). - Yeast asparagine/glutamine permease (gene AGP1). - Yeast glutamine permease (gene GNP1). - Yeast histidine permease (gene HIP1). - Yeast lysine permease (gene LYP1). - Yeast proline permease (gene PUT4). - Yeast valine and tyrosine permease (gene VAL1/TAT1). - Yeast tryptophan permease (gene TAT2/SCM2). - Yeast choline transport protein (gene HNM1/CTR1). - Yeast GABA permease (gene UGA4). - Yeast hypothetical protein YKL174c. - Fission yeast protein isp5. - Fission yeast hypothetical protein SpAC8A4.11 - Fission yeast hypothetical protein SpAC11D3.08c. - Emericella nidulans proline transport protein (gene prnB). - Trichoderma harzianum amino acid permease INDA1. - Salmonella typhimurium L-asparagine permease (gene ansP). - Escherichia coli aromatic amino acid transport protein (gene aroP). - Escherichia coli D-serine/D-alanine/glycine transporter (gene cycA). - Escherichia coli GABA permease (gene gabP). - Escherichia coli lysine-specific permease (gene lysP). - Escherichia coli phenylalanine-specific permease (gene pheP). - Salmonella typhimurium proline-specific permease (gene proY). - Escherichia coli and Klebsiella pneumoniae hypothetical protein yeeF. - Escherichia coli and Salmonella typhimurium hypothetical protein yifK. - Bacillus subtilis permeases rocC and rocE which probably transports arginine or ornithine. These proteins seem to contain up to 12 transmembrane segments. As a signature for this family of proteins, the best conserved region which is located in the second transmembrane segment has been selected.

Consensus pattern: ~~[STAGC]~~[STAGC (SEQ ID NO: 691)]-G-[PAG]-x(2,3)-~~[LIVMFYWA]~~[LIVMFYWA (SEQ ID NO: 464)](2)-x-~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)]-x- ~~[LIVMFWSTAGC]~~[LIVMFWSTAGC (SEQ ID NO: 433)](2)-~~[STAGC]~~[STAGC (SEQ ID NO: 691)]-x(3)-~~[LIVMFYWT]~~[LIVMFYWT (SEQ ID NO: 485)]-x-~~[LIVMST]~~[LIVMST (SEQ ID NO: 509)]-x(3)- ~~[LIVMCTA]~~[LIVMCTA (SEQ ID NO: 399)]-[GA]-E-x(5)-~~[PSAL]~~[PSAL (SEQ ID NO: 611)]-

[ 1] Weber E., Chevalier M.R., Jund R. J. Mol. Evol. 27:341-350(1988).

[ 2] Vandenbol M., Jauniaux J.-C., Grenson M. Gene 83:153-159(1989).

[ 3] Reizer J., Finley K., Kakuda D., McLeod C.L., Reizer A., Saier M.H. Jr. Protein Sci. 2:20-30(1993).

38. aakinase (1) Glutamate 5-kinase signature

Glutamate 5-kinase (EC 2.7.2.11) (gamma-glutamyl kinase) (GK) is the enzyme that catalyzes the first step in the biosynthesis of proline from glutamate, the ATP-dependent phosphorylation of L-glutamate into L-glutamate 5-phosphate. In eubacteria (gene proB) and yeast [1] (gene PRO1), GK is a monofunctional protein, while in plants and mammals, it is a bifunctional enzyme (P5CS) [2]that consists of two domains: a N-terminal GK domain and a C-terminal gamma-glutamyl phosphate reductase domain (EC 1.2.1.41) (see <PDOC00940>).As a signature pattern, a highly conserved glycine-and alanine-rich region located in the central section of these enzymes has been selected. Yeast hypothetical protein YHR033w is highly similar to GK.

Consensus pattern: ~~[GSTN]~~[GSTN (SEQ ID NO: 244)]-x(2)-G-x-G-[GC]-[IM]-x-[STA]-K-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-[SA]-[TCA]- x(2)-~~[GALV]~~[GALV (SEQ ID NO: 169)]-x(3)-G-

[ 1] Li W., Brandriss M.C. J. Bacteriol. 174:4148-4156(1992).
[ 2] Hu C.-A.A., Delauney A.J., Verma D.P.S. Proc. Natl. Acad. Sci. U.S.A. 89:9354-9358(1992).

aakinase (2) Aspartokinase signature

Aspartokinase (EC 2.7.2.4) (AK) [1] catalyzes the phosphorylation of aspartate. The product of this reaction can then be used in the biosynthesis of lysine or in the pathway leading to homoserine, which participates in the biosynthesis of threonine, isoleucine and methionine. In Escherichia coli, there are three different isozymes which differ in their sensitivity to repression and inhibition by Lys, Met and Thr. AK1 (gene thrA) and AK2 (gene metL) are bifunctional enzymes which both consist of an N- terminal AK domain and a C-terminal homoserine dehydrogenase domain. AK1 is involved in threonine biosynthesis and AK2, in that of methionine. The third isozyme, AK3 (gene lysC), is monofunctional and involved in lysine synthesis. In yeast, there is a single isozyme of AK (gene HOM3). As a signature pattern for AK, a conserved region located in the N-terminal extremity has been selected.

Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-K-[FY]-G-G-[ST]-[SC]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-

[ 1] Rafalski J.A., Falco S.C. J. Biol. Chem. 263:2146-2151(1988).

aakinase (3) Gamma-glutamyl phosphate reductase signature
Gamma-glutamyl phosphate reductase (EC 1.2.1.41) (GPR) is the enzyme that catalyzes the second step in the biosynthesis of proline from glutamate, the NADP-dependent reduction of L-glutamate 5-phosphate into L-glutamate 5-semialdehyde and phosphate. In eubacteria (gene proA) and yeast [1] (gene PRO2), GPR is a monofunctional protein, while in plants and mammals, it is a bifunctional enzyme (P5CS) [2]that consists of two domains: a N-terminal glutamate 5-kinase domain(EC 2.7.2.11) (see <PDOC00701>) and a C-terminal GPR domain. As a signature pattern, a conserved region that contains two histidine residues has been selected. This region is located in the last third of GPR.

Consensus pattern: V-x(5)-A-[LIV]-x-H-I-x(2)-[HY]-[GS]-[ST]-x-H-[ST]-[DE]-x- I-

[ 1] Pearson B.M., Hernando Y., Payne J., Wolf S.S., Kalogeropoulos A., Schweizer M. Yeast 12:1021-1031(1996).
[ 2] Hu C.-A.A., Delauney A.J., Verma D.P.S. Proc. Natl. Acad. Sci. U.S.A. 89:9354-9358(1992).

39. (abhydrolase) alpha/beta hydrolase fold. This catalytic domain is found in a very wide range of enzymes.

[1] Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschueren KHG, Goldman A, Protein Eng 1992;5:197-211.

40. (Acid phosphat) Histidine acid phosphatases signatures
Acid phosphatases (EC 3.1.3.2) are a heterogeneous group of proteins that hydrolyze phosphate esters, optimally at low pH. It has been shown [1] that a number of acid phosphatases, from both prokaryotes and eukaryotes, share two regions of sequence similarity, each centered around a conserved histidine residue. These two histidines seem to be involved in the enzymes' catalytic mechanism [2,3]. The first histidine is located in the N-terminal section and forms a phosphohistidine intermediate while the second is located in

the C- terminal section and possibly acts as proton donor. Enzymes belonging to this family are called 'histidine acid phosphatases' and are listed below:

- Escherichia coli pH 2.5 acid phosphatase (gene appA).

- Escherichia coli glucose-1-phosphatase (EC 3.1.3.10) (gene agp).
- Yeast constitutive and repressible acid phosphatases (genes PHO3 and PHO5).
- Fission yeast acid phosphatase (gene pho1).
- Aspergillus phytases A and B (EC 3.1.3.8) (gene phyA and phyB).
- Mammalian lysosomal acid phosphatase.
- Mammalian prostatic acid phosphatase.
- Caenorhabditis elegans hypothetical proteins B0361.7, C05C10.1, C05C10.4 and F26C11.1.

Consensus pattern~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(2)-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)]-x(2)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-R-H-[GN]-x-R-x-[PAS] [H is the phosphohistidine residue]

Consensus pattern~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x-~~[LIVMFAG]~~[LIVMFAG (SEQ ID NO: 405)]-x(2)-~~[STAGI]~~[STAGI (SEQ ID NO: 700)]-H-D-~~[STANQ]~~[STANQ (SEQ ID NO: 724)]-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(2)-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x(2)-[STA] [H is an active site residue] Sequences known to belong to this class detected by the patternALL, except for rat prostatic acid phosphatase which seems to have Tyr instead of the active site His

[ 1] van Etten R.L., Davidson R., Stevis P.E., MacArthur H., Moore D.L. J. Biol. Chem. 266:2313-2319(1991).

[ 2] Ostanin K., Harms E.H., Stevis P.E., Kuciel R., Zhou M.-M., van Etten R.L. J. Biol. Chem. 267:22830-22836(1992).

[ 3] Schneider G., Lindqvist Y., Vihko P. EMBO J. 12:2609-2615(1993).

41. Aconitase family signatures

Aconitase (aconitate hydratase) (EC 4.2.1.3) [1] is the enzyme from the tricarboxylic acid cycle that catalyzes the reversible isomerization of citrate and isocitrate. Cis-aconitate is

formed as an intermediary product during the course of the reaction. In eukaryotes two isozymes of aconitase are known to exist: one found in the mitochondrial matrix and the other found in the cytoplasm. Aconitase, in its active form, contains a 4Fe-4S iron-sulfur cluster; three cysteine residues have been shown to be ligands of the 4Fe-4S cluster. It has been shown that the aconitase family also contains the following proteins: - Iron-responsive element binding protein (IRE-BP). IRE-BP is a cytosolic protein that binds to iron-responsive elements (IREs). IREs are stem-loop structures found in the 5'UTR of ferritin, and delta aminolevulinic acid synthase mRNAs, and in the 3'UTR of transferrin receptor mRNA. IRE-BP also express aconitase activity. - 3-isopropylmalate dehydratase (EC 4.2.1.33) (isopropylmalate isomerase), the enzyme that catalyzes the second step in the biosynthesis of leucine. - Homoaconitase (EC 4.2.1.36) (homoaconitate hydratase), an enzyme that participates in the alpha-aminoadipate pathway of lysine biosynthesis and that converts cis-homoaconitate into homoisocitric acid. - Escherichia coli protein ybhJ

Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(2)-~~[GSACIVM]~~[GSACIVM (SEQ ID NO: 190)]-x-[LIV]-~~[GTIV]~~[GTIV (SEQ ID NO: 255)]-[STP]-C-x(0,1)-T-N-~~[GSTANI]~~[GSTANI (SEQ ID NO: 235)]-x(4)-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)] [C binds the iron-sulfur center]

Consensus pattern: G-x(2)-~~[LIVWPQ]~~[LIVWPQ (SEQ ID NO: 539)]-x(3)-[GAC]-C-~~[GSTAM]~~[GSTAM (SEQ ID NO: 232)]-~~[LIMPTA]~~[LIMPTA (SEQ ID NO: 343)]-C-~~[LIMV]~~[LIMV (SEQ ID NO: 347)]- [GA] [The two C's bind the iron-sulfur center]-

[ 1] Gruer M.J., Artymiuk P.J., Guest J.R. Trends Biochem. Sci. 22:3-6(1997).

42. Actins signatures

Actins [1 to 4] are highly conserved contractile proteins that are present in all eukaryotic cells. In vertebrates there are three groups of actin isoforms: alpha, beta and gamma. The alpha actins are found in muscle tissues and are a major constituent of the contractile apparatus. The beta and gamma actins co-exists in most cell types as components of the cytoskeleton and as mediators of internal cell motility. In plants [5] there are many isoforms which are probably involved in a variety of functions such as cytoplasmic streaming, cell shape determination, tip growth, graviperception, cell wall deposition, etc. Actin exists either in a monomeric form (G-actin) or in a polymerized form (F-actin). Each actin monomer can

bind a molecule of ATP; when polymerization occurs, the ATP is hydrolyzed. Actin is a protein of from 374 to 379 amino acid residues. The structure of actin has been highly conserved in the course of evolution. Recently some divergent actin-like proteins have been identified in several species. These proteins are: - Centractin (actin-RPV) from mammals, fungi (yeast ACT5, Neurospora crassa ro-4) and Pneumocystis carinii (actin-II). Centractin seems to be a component of a multi-subunit centrosomal complex involved in microtubule based vesicle motility. This subfamily is also known as ARP1. - ARP2 subfamily which includes chicken ACTL, yeast ACT2, Drosophila 14D, C.elegans actC. - ARP3 subfamily which includes actin 2 from mammals, Drosophila 66B, yeast ACT4 and fission yeast act2. - ARP4 subfamily which includes yeast ACT3 and Drosophila 13E. Three signature patterns have been developed. The first two are specific to actins and span positions 54 to 64 and 357 to 365. The last signature picks up both actins and the actin-like proteins and corresponds to positions 106 to 118 in actins.

Consensus pattern: [FY]-[LIV]-G-[DE]-E-A-Q-x-[RKQ](2)-G-
Consensus pattern: W-[IV]-[STA]-[RK]-x-[DE]-Y-[DNE]-[DE]-
Consensus pattern: [LM]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-T-E-~~[GAPQ]~~[GAPQ (SEQ ID NO: 171)]-x-~~[LIVMFYWHQ]~~[LIVMFYWHQ (SEQ ID NO: 476)]-N-~~[PSTAQ]~~[PSTAQ (SEQ ID NO: 618)]-x(2)-N-[KR]-

[ 1] Sheterline P., Clayton J., Sparrow J.C. (In) Actins, 3rd Edition, Academic Press Ltd, London, (1996).
[ 2] Pollard T.D., Cooper J.A. Annu. Rev. Biochem. 55:987-1036(1986).
[ 3] Pollard T.D. Curr. Opin. Cell Biol. 1:33-40(1990).
[ 4] Rubenstein P.A. BioEssays 12:309-315(1990).
[ 5] Meagher R.B., McLean B.G. Cell Motil. Cytoskeleton 16:164-166(1990).

43. Adenylate kinase signature

Adenylate kinase (EC 2.7.4.3) (AK) [1] is a small monomeric enzyme that catalyzes the reversible transfer of MgATP to AMP (MgATP + AMP = MgADP + ADP).In mammals there are three different isozymes: - AK1 (or myokinase), which is cytosolic. - AK2, which is located in the outer compartment of mitochondria. - AK3 (or GTP:AMP phosphotransferase), which is located in the mitochondrial matrix and which uses MgGTP instead of MgATP.The

sequence of AK has also been obtained from different bacterial species and from plants and fungi. Two other enzymes have been found to be evolutionary related to AK. These are: - Yeast uridylate kinase (EC 2.7.4.-) (UK) (gene URA6) [2] which catalyzes the transfer of a phosphate group from ATP to UMP to form UDP and ADP. - Slime mold UMP-CMP kinase (EC 2.7.4.14) [3] which catalyzes the transfer of a phosphate group from ATP to either CMP or UMP to form CDP or UDP and ADP. Several regions of AK family enzymes are well conserved, including the ATP-binding domains. The most conserved of all regions have been selected as a signature for this type of enzyme. This region includes an aspartic acid residue that is part of the catalytic cleft of the enzyme and that is involved in a salt bridge. It also includes an arginine residue whose modification leads to inactivation of the enzyme

Consensus pattern: ~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)](3)-D-G-[FYI]-P-R-x(3)-[NQ]-

[ 1] Schulz G.E. Cold Spring Harbor Symp. Quant. Biol. 52:429-439(1987).
[ 2] Liljelund P., Sanni A., Friesen J.D., Lacroute F. Biochem. Biophys. Res. Commun. 165:464-473(1989).
[ 3] Wiesmueller L., Noegel A.A., Barzu O., Gerisch G., Schleicher M. J. Biol. Chem. 265:6339-6345(1990).
[ 4] Kath T.H., Schmid R., Schaefer G. Arch. Biochem. Biophys. 307:405-410(1993).

44. (adh_short) Short-chain dehydrogenases/reductases family signature. The short-chain dehydrogenases/reductases family (SDR) [1] is a very large family of enzymes, most of which are known to be NAD- or NADP-dependent oxidoreductases. As the first member of this family to be characterized was Drosophila alcohol dehydrogenase, this family used to be called [2,3,4]'insect-type', or 'short-chain' alcohol dehydrogenases. Most member of this family are proteins of about 250 to 300 amino acid residues. The proteins currently known to belong to this family are listed below. - Alcohol dehydrogenase (EC 1.1.1.1) from insects such as Drosophila. - Acetoin dehydrogenase (EC 1.1.1.5) from Klebsiella terrigena (gene budC). - D-beta-hydroxybutyrate dehydrogenase (BDH) (EC 1.1.1.30) from mammals. - Acetoacetyl-CoA reductase (EC 1.1.1.36) from various bacterial species (gene phbB or phaB). - Glucose 1-dehydrogenase (EC 1.1.1.47) from Bacillus. - 3-beta-hydroxysteroid dehydrogenase (EC 1.1.1.51) from Comomonas testosteroni. - 20-beta-hydroxysteroid

dehydrogenase (EC 1.1.1.53) from Streptomyces hydrogenans. - Ribitol dehydrogenase (EC 1.1.1.56) (RDH) from Klebsiella aerogenes. - Estradiol 17-beta-dehydrogenase (EC 1.1.1.62) from human. - Gluconate 5-dehydrogenase (EC 1.1.1.69) from Gluconobacter oxydans (gene gno). - 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) from Escherichia coli (gene fabG) and from plants. - Retinol dehydrogenase (EC 1.1.1.105) from mammals. - 2-deoxy-d-gluconate 3-dehydrogenase (EC 1.1.1.125) from Escherichia coli and Erwinia chrysanthemi (gene kduD). - Sorbitol-6-phosphate 2-dehydrogenase (EC 1.1.1.140) from Escherichia coli (gene gutD) and from Klebsiella pneumoniae (gene sorD). - 15-hydroxyprostaglandin dehydrogenase (NAD+) (EC 1.1.1.141) from human. - Corticosteroid 11-beta-dehydrogenase (EC 1.1.1.146) (11-DH) from mammals. - 7-alpha-hydroxysteroid dehydrogenase (EC 1.1.1.159) from Escherichia coli (gene hdhA), Eubacterium strain VPI 12708 (gene baiA) and from Clostridium sordellii. - NADPH-dependent carbonyl reductase (EC 1.1.1.184) from mammals. - Tropinone reductase-I (EC 1.1.1.206) and -II (EC 1.1.1.236) from plants. - N-acylmannosamine 1-dehydrogenase (EC 1.1.1.233) from Flavobacterium strain 141-8. - D-arabinitol 2-dehydrogenase (ribulose forming) (EC 1.1.1.250) from fungi. - Tetrahydroxynaphthalene reductase (EC 1.1.1.252) from Magnaporthe grisea. - Pteridine reductase 1 (EC 1.1.1.253) (gene PTR1) from Leishmania. - 2,5-dichloro-2,5-cyclohexadiene-1,4-diol dehydrogenase (EC 1.1.-.-) from Pseudomonas paucimobilis. - Cis-1,2-dihydroxy-3,4-cyclohexadiene-1-carboxylate dehydrogenase (EC 1.3.1. -) from Acinetobacter calcoaceticus (gene benD) and Pseudomonas putida (gene xylL). - Biphenyl-2,3-dihydro-2,3-diol dehydrogenase (EC 1.3.1.-) (gene bphB) from various Pseudomonaceae. - Cis-toluene dihydrodiol dehydrogenase (EC 1.3.1.-) from Pseudomonas putida (gene todD). - Cis-benzene glycol dehydrogenase (EC 1.3.1.19) from Pseudomonas putida (gene bnzE). - 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (EC 1.3.1.28) from Escherichia coli (gene entA) and Bacillus subtilis (gene dhbA). - Dihydropteridine reductase (EC 1.6.99.7) (HDHPR) from mammals. - Lignin degradation enzyme ligD from Pseudomonas paucimobilis. - Agropine synthesis reductase from Agrobacterium plasmids (gene mas1). - Versicolorin reductase from Aspergillus parasiticus (gene VER1). - Putative keto-acyl reductases from Streptomyces polyketide biosynthesis operons. - A trifunctional hydratase-dehydrogenase-epimerase from the peroxisomal beta-oxidation system of Candida tropicalis. This protein contains two tandemly repeated 'short-chain dehydrogenase-type' domain in its N-terminal extremity. - Nodulation protein nodG from species of Azospirillum and Rhizobium which is probably involved in the modification of the nodulation Nod factor fatty acyl chain. - Nitrogen fixation protein fixR from Bradyrhizobium japonicum. - Bacillus

subtilis protein dltE which is involved in the biosynthesis of D- alanyl-lipoteichoic acid. - Human follicular variant translocation protein 1 (FVT1). - Mouse adipocyte protein p27. - Mouse protein Ke 6. - Maize sex determination protein TASSELSEED 2. - Sarcophaga peregrina 25 Kd development specific protein. - Drosophila fat body protein P6. - A Listeria monocytogenes hypothetical protein encoded in the internalins gene region. - Escherichia coli hypothetical protein yciK. - Escherichia coli hypothetical protein ydfG. - Escherichia coli hypothetical protein yjgI. - Escherichia coli hypothetical protein yjgU. - Escherichia coli hypothetical protein yohF. - Bacillus subtilis hypothetical protein yoxD. - Bacillus subtilis hypothetical protein ywfD. - Bacillus subtilis hypothetical protein ywfH. - Yeast hypothetical protein YIL124w. - Yeast hypothetical protein YIR035c. - Yeast hypothetical protein YIR036c. - Yeast hypothetical protein YKL055c. - Fission yeast hypothetical protein SpAC23D3.11. One of the best conserved regions which includes two perfectly conserved residues, a tyrosine and a lysine has been selected as a signature pattern for this family of proteins. The tyrosine residue participates in the catalytic mechanism.

Consensus pattern: ~~[LIVSPADNK]~~[LIVSPADNK (SEQ ID NO: 535)]-x(12)-Y-~~[PSTAGNCV]~~[PSTAGNCV (SEQ ID NO: 617)]-~~[STAGNQCIVM]~~[STAGNQCIVM (SEQ ID NO: 706)]-~~[STAGC]~~[STAGC (SEQ ID NO: 691)]-K- {PC}-~~[SAGFYR]~~[SAGFYR (SEQ ID NO: 659)]-~~[LIVMSTAGD]~~[LIVMSTAGD (SEQ ID NO: 516)]-x(2)-~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)]-x(3)-~~[LIVMFYWGAPTHQ]~~[LIVMFYWGAPTHQ (SEQ ID NO: 472)]-~~[GSACQRHM]~~[GSACQRHM (SEQ ID NO: 193)] [Y is an active site residue] -

[ 1] Joernvall H., Persson B., Krook M., Atrian S., Gonzalez-Duarte R., Jeffery J., Ghosh D. Biochemistry 34:6003-6013(1995).

[ 2] Villarroya A., Juan E., Egestad B., Joernvall H. Eur. J. Biochem. 180:191-197(1989).

[ 3] Persson B., Krook M., Joernvall H. Eur. J. Biochem. 200:537-543(1991).

[ 4] Neidle E.L., Hartnett C., Ornston N.L., Bairoch A., Rekik M., Harayama S. Eur. J. Biochem. 204:113-120(1992).

45. (adh_short_C2) Short-chain dehydrogenases/reductases family signature

The short-chain dehydrogenases/reductases family (SDR) [1] is a very large family of enzymes, most of which are known to be NAD- or NADP-dependent oxidoreductases. As the

first member of this family to be characterized was Drosophila alcohol dehydrogenase, this family used to be called [2,3,4]'insect-type', or 'short-chain' alcohol dehydrogenases. Most member of this family are proteins of about 250 to 300 amino acid residues. The proteins currently known to belong to this family are listed below. - Alcohol dehydrogenase (EC 1.1.1.1) from insects such as Drosophila. - Acetoin dehydrogenase (EC 1.1.1.5) from Klebsiella terrigena (gene budC). - D-beta-hydroxybutyrate dehydrogenase (BDH) (EC 1.1.1.30) from mammals. - Acetoacetyl-CoA reductase (EC 1.1.1.36) from various bacterial species (gene phbB or phaB). - Glucose 1-dehydrogenase (EC 1.1.1.47) from Bacillus. - 3-beta-hydroxysteroid dehydrogenase (EC 1.1.1.51) from Comomonas testosteroni. - 20-beta-hydroxysteroid dehydrogenase (EC 1.1.1.53) from Streptomyces hydrogenans. - Ribitol dehydrogenase (EC 1.1.1.56) (RDH) from Klebsiella aerogenes. - Estradiol 17-beta-dehydrogenase (EC 1.1.1.62) from human. - Gluconate 5-dehydrogenase (EC 1.1.1.69) from Gluconobacter oxydans (gene gno). - 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) from Escherichia coli (gene fabG) and from plants. - Retinol dehydrogenase (EC 1.1.1.105) from mammals. - 2-deoxy-d-gluconate 3-dehydrogenase (EC 1.1.1.125) from Escherichia coli and Erwinia chrysanthemi (gene kduD). - Sorbitol-6-phosphate 2-dehydrogenase (EC 1.1.1.140) from Escherichia coli (gene gutD) and from Klebsiella pneumoniae (gene sorD). - 15-hydroxyprostaglandin dehydrogenase (NAD+) (EC 1.1.1.141) from human. - Corticosteroid 11-beta-dehydrogenase (EC 1.1.1.146) (11-DH) from mammals. - 7-alpha-hydroxysteroid dehydrogenase (EC 1.1.1.159) from Escherichia coli (gene hdhA), Eubacterium strain VPI 12708 (gene baiA) and from Clostridium sordellii. - NADPH-dependent carbonyl reductase (EC 1.1.1.184) from mammals. - Tropinone reductase-I (EC 1.1.1.206) and -II (EC 1.1.1.236) from plants. - N-acylmannosamine 1-dehydrogenase (EC 1.1.1.233) from Flavobacterium strain 141-8. - D-arabinitol 2-dehydrogenase (ribulose forming) (EC 1.1.1.250) from fungi. - Tetrahydroxynaphthalene reductase (EC 1.1.1.252) from Magnaporthe grisea. - Pteridine reductase 1 (EC 1.1.1.253) (gene PTR1) from Leishmania. - 2,5-dichloro-2,5-cyclohexadiene-1,4-diol dehydrogenase (EC 1.1.-.-) from Pseudomonas paucimobilis. - Cis-1,2-dihydroxy-3,4-cyclohexadiene-1-carboxylate dehydrogenase (EC 1.3.1. -) from Acinetobacter calcoaceticus (gene benD) and Pseudomonas putida (gene xylL). - Biphenyl-2,3-dihydro-2,3-diol dehydrogenase (EC 1.3.1.-) (gene bphB) from various Pseudomonaceae. - Cis-toluene dihydrodiol dehydrogenase (EC 1.3.1.-) from Pseudomonas putida (gene todD). - Cis-benzene glycol dehydrogenase (EC 1.3.1.19) from Pseudomonas putida (gene bnzE). - 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (EC 1.3.1.28) from Escherichia coli (gene entA) and Bacillus subtilis (gene

dhbA). - Dihydropteridine reductase (EC 1.6.99.7) (HDHPR) from mammals. - Lignin degradation enzyme ligD from Pseudomonas paucimobilis. - Agropine synthesis reductase from Agrobacterium plasmids (gene mas1). - Versicolorin reductase from Aspergillus parasiticus (gene VER1). - Putative keto-acyl reductases from Streptomyces polyketide biosynthesis operons. - A trifunctional hydratase-dehydrogenase-epimerase from the peroxisomal beta-oxidation system of Candida tropicalis. This protein contains two tandemly repeated 'short-chain dehydrogenase-type' domain in its N-terminal extremity. - Nodulation protein nodG from species of Azospirillum and Rhizobium which is probably involved in the modification of the nodulation Nod factor fatty acyl chain. - Nitrogen fixation protein fixR from Bradyrhizobium japonicum. - Bacillus subtilis protein dltE which is involved in the biosynthesis of D- alanyl-lipoteichoic acid. - Human follicular variant translocation protein 1 (FVT1). - Mouse adipocyte protein p27. - Mouse protein Ke 6. - Maize sex determination protein TASSELSEED 2. - Sarcophaga peregrina 25 Kd development specific protein. - Drosophila fat body protein P6. - A Listeria monocytogenes hypothetical protein encoded in the internalins gene region. - Escherichia coli hypothetical protein yciK. - Escherichia coli hypothetical protein ydfG. - Escherichia coli hypothetical protein yjgI. - Escherichia coli hypothetical protein yjgU. - Escherichia coli hypothetical protein yohF. - Bacillus subtilis hypothetical protein yoxD. - Bacillus subtilis hypothetical protein ywfD. - Bacillus subtilis hypothetical protein ywfH. - Yeast hypothetical protein YIL124w. - Yeast hypothetical protein YIR035c. - Yeast hypothetical protein YIR036c. - Yeast hypothetical protein YKL055c. - Fission yeast hypothetical protein SpAC23D3.11. One of the best conserved regions which includes two perfectly conserved residues, a tyrosine and a lysine has been used as a signature pattern for this family of proteins. The tyrosine residue participates in the catalytic mechanism.

Consensus pattern: ~~[LIVSPADNK]~~[LIVSPADNK (SEQ ID NO: 535)]-x(12)-Y-~~[PSTAGNCV]~~[PSTAGNCV (SEQ ID NO: 617)]-~~[STAGNQCIVM]~~[STAGNQCIVM (SEQ ID NO: 706)]-~~[STAGC]~~[STAGC (SEQ ID NO: 691)]-K- {PC}-~~[SAGFYR]~~[SAGFYR (SEQ ID NO: 659)]-~~[LIVMSTAGD]~~[LIVMSTAGD (SEQ ID NO: 516)]-x(2)-~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)]-x(3)-~~[LIVMFYWGAPTHQ]~~[LIVMFYWGAPTHQ (SEQ ID NO: 472)]-~~[GSACQRHM]~~[GSACQRHM (SEQ ID NO: 193)] [Y is an active site residue]

[ 1] Joernvall H., Persson B., Krook M., Atrian S., Gonzalez-Duarte R., Jeffery J., Ghosh D. Biochemistry 34:6003-6013(1995).

[ 2] Villarroya A., Juan E., Egestad B., Joernvall H. Eur. J. Biochem. 180:191-197(1989).

[ 3] Persson B., Krook M., Joernvall H. Eur. J. Biochem. 200:537-543(1991).

[ 4] Neidle E.L., Hartnett C., Ornston N.L., Bairoch A., Rekik M., Harayama S. Eur. J. Biochem. 204:113-120(1992).

46. (adh_zinc) Zinc-containing alcohol dehydrogenases signatures

Alcohol dehydrogenase (EC 1.1.1.1) (ADH) catalyzes the reversible oxidation of ethanol to acetaldehyde with the concomitant reduction of NAD [1]. Currently three, structurally and catalytically, different types of alcohol dehydrogenases are known: - Zinc-containing 'long-chain' alcohol dehydrogenases. - Insect-type, or 'short-chain' alcohol dehydrogenases. - Iron-containing alcohol dehydrogenases.Zinc-containing ADH's [2,3] are dimeric or tetrameric enzymes that bind two atoms of zinc per subunit. One of the zinc atom is essential for catalytic activity while the other is not. Both zinc atoms are coordinated by either cysteine or histidine residues; the catalytic zinc is coordinated by two cysteines and one histidine. Zinc-containing ADH's are found in bacteria, mammals, plants, and in fungi. In most species there are more than one isozyme (for example, human have at least six isozymes, yeast have three, etc.). A number of other zinc-dependent dehydrogenases are closely related to zinc ADH [4], these are: - Xylitol dehydrogenase (EC 1.1.1.9) (D-xylulose reductase). - Sorbitol dehydrogenase (EC 1.1.1.14). - Aryl-alcohol dehydrogenase (EC 1.1.1.90) (benzyl alcohol dehydrogenase). - Threonine 3-dehydrogenase (EC 1.1.1.103). - Cinnamyl-alcohol dehydrogenase (EC 1.1.1.195) (CAD) [5]. CAD is a plant enzyme involved in the biosynthesis of lignin. - Galactitol-1-phosphate dehydrogenase (EC 1.1.1.251). - Pseudomonas putida 5-exo-alcohol dehydrogenase (EC 1.1.1.-) [6]. - Escherichia coli starvation sensing protein rspB. - Escherichia coli hypothetical protein yjgB. - Escherichia coli hypothetical protein yjgV. - Escherichia coli hypothetical protein yjjN. - Yeast hypothetical protein YAL060w (FUN49). - Yeast hypothetical protein YAL061w (FUN50). - Yeast hypothetical protein YCR105w. The pattern that has been developed to detect this class of enzymes is based on a conserved region that includes a histidine residue which is the second ligand of the catalytic zinc atom. This family also includes NADP-dependent quinone oxidoreductase (EC 1.6.5.5),an enzyme found in bacteria (gene qor), in yeast and in mammals where, in some species such as rodents, it has been recruited as an eye lens protein

and is known as zeta-crystallin [7]. The sequence of quinone oxidoreductase is distantly related to that other zinc-containing alcohol dehydrogenases and it lacks the zinc-ligand residues. The torpedo fish and mammlian synaptic vesicle membrane protein vat-1 is related to qor. A specific pattern has been developed for this subfamily.

Consensus pattern: G-H-E-x(2)-G-x(5)-[GA]-x(2)-~~[IVSAC]~~[IVSAC (SEQ ID NO: 284)] [H is a zinc ligand]

Consensus pattern: [GSD]-~~[DEQH]~~[DEQH (SEQ ID NO: 69)]-x(2)-L-x(3)-[SA](2)-G-G-x-G-x(4)-Q-x(2)-[KR]-

[ 1] Branden C.-I., Joernvall H., Eklund H., Furugren B. (In) The Enzymes (3rd edition) 11:104-190(1975).

[ 2] Joernvall H., Persson B., Jeffery J. Eur. J. Biochem. 167:195-201(1987).

[ 3] Sun H.-W., Plapp B.V. J. Mol. Evol. 34:522-535(1992).

[ 4] Persson B., Hallborn J., Walfridsson M., Hahn-Haegerdal B., Keraenen S., Penttilae M., Joernvall H. FEBS Lett. 324:9-14(1993).

[ 5] Knight M.E., Halpin C., Schuch W. Plant Mol. Biol. 19:793-801(1992).

[ 6] Koga H., Aramaki H., Yamaguchi E., Takeuchi K., Horiuchi T., Gunsalus I.C. J. Bacteriol. 166:1089-1095(1986).

[ 7] Joernvall H., Persson B., Du Bois G., Lavers G.C., Chen J.H., Gonzalez P., Rao P.V., Zigler J.S. Jr. FEBS Lett. 322:240-244(1993).

47. (aldedh) Aldehyde dehydrogenases active sites

Aldehyde dehydrogenases (EC 1.2.1.3 and EC 1.2.1.5) are enzymes which oxidize a wide variety of aliphatic and aromatic aldehydes. In mammals at least four different forms of the enzyme are known [1]: class-1 (or Ald C) a tetrameric cytosolic enzyme, class-2 (or Ald M) a tetrameric mitochondrial enzyme, class-3 (or Ald D) a dimeric cytosolic enzyme, and class IV a microsomal enzyme. Aldehyde dehydrogenases have also been sequenced from fungal and bacterial species. A number of enzymes are known to be evolutionary related to aldehyde dehydrogenases; these enzymes are listed below. - Plants and bacterial betaine-aldehyde dehydrogenase (EC 1.2.1.8) [2], an enzyme that catalyzes the last step in the biosynthesis of betaine. - Plants and bacterial NADP-dependent glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.9). - Escherichia coli succinate-semialdehyde dehydrogenase (NADP+) (EC

1.2.1.16) (gene gabD) [3], which reduces succinate semialdehyde into succinate. - Escherichia coli lactaldehyde dehydrogenase (EC 1.2.1.22) (gene ald) [4]. - Mammalian succinate semialdehyde dehydrogenase (NAD+) (EC 1.2.1.24). - Escherichia coli phenylacetaldehyde dehydrogenase (EC 1.2.1.39). - Escherichia coli 5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (gene hpcC). - Pseudomonas putida 2-hydroxymuconic semialdehyde dehydrogenase [5] (genes dmpC and xylG), an enzyme in the meta-cleavage pathway for the degradation of phenols, cresols and catechol. - Bacterial and mammalian methylmalonate-semialdehyde dehydrogenase (MMSDH) (EC 1.2.1.27) [6], an enzyme involved in the distal pathway of valine catabolism. - Yeast delta-1-pyrroline-5-carboxylate dehydrogenase (EC 1.5.1.12) [7] (gene PUT2), which converts proline to glutamate. - Bacterial multifunctional putA protein, which contains a delta-1-pyrroline- 5-carboxylate dehydrogenase domain. - 26G, a garden pea protein of unknown function which is induced by dehydration of shoots [8]. - Mammalian formyltetrahydrofolate dehydrogenase (EC 1.5.1.6) [9]. This is a cytosolic enzyme responsible for the NADP-dependent decarboxylative reduction of 10-formyltetrahydrofolate into tetrahydrofolate. It is an protein of about 900 amino acids which consist of three domains; the C- terminal domain (480 residues) is structurally and functionally related to aldehyde dehydrogenases. - Yeast hypothetical protein YBR006w. - Yeast hypothetical protein YER073w. - Yeast hypothetical protein YHR039c. - Caenorhabditis elegans hypothetical protein F01F1.6.A glutamic acid and a cysteine residue have been implicated in the catalytic activity of mammalian aldehyde dehydrogenase. These residues are conserved in all the enzymes of this family. Two patterns have been derived for this family, one for each of the active site residues.

Consensus pattern: ~~[LIVMFGA]~~[LIVMFGA (SEQ ID NO: 415)]-E-~~[LIMSTAC]~~[LIMSTAC (SEQ ID NO: 345)]-[GS]-G-~~[KNLM]~~[KNLM (SEQ ID NO: 285)]-~~[SADN]~~[SADN (SEQ ID NO: 655)]-~~[TAPFV]~~[TAPFV (SEQ ID NO: 763)] [E is the active site residue]-
Consensus pattern: ~~[FYLVA]~~[FYLVA (SEQ ID NO: 131)]-x(3)-G-[QE]-x-C-~~[LIVMGSTANC]~~[LIVMGSTANC (SEQ ID NO: 493)]-~~[AGCN]~~[AGCN (SEQ ID NO: 1)]-x-~~[GSTADNEKR]~~[GSTADNEKR (SEQ ID NO: 223)] [C is the active site residue

[ 1] Hempel J., Harper K., Lindahl R. Biochemistry 28:1160-1167(1989).

[ 2] Weretilnyk E.A., Hanson A.D. Proc. Natl. Acad. Sci. U.S.A. 87:2745-2749(1990).

[ 3] Niegemann E., Schulz A., Bartsch K. Arch. Microbiol. 160:454-460(1993).

[ 4] Hidalgo E., Chen Y.-M., Lin E.C.C., Aguilar J. J. Bacteriol. 173:6118-6123(1991).

[ 5] Nordlund I., Shingler V. Biochim. Biophys. Acta 1049:227-230(1990).

[ 6] Steele M.I., Lorenz D., Hatter K., Park A., Sokatch J.R. J. Biol. Chem. 267:13585-13592(1992).

[ 7] Krzywicki K.A., Brandriss M.C. Mol. Cell. Biol. 4:2837-2842(1984).

[ 8] Guerrero F.D., Jones J.T., Mullet J.E. Plant Mol. Biol. 15:11-26(1990).

[ 9] Cook R.J., Lloyd R.S., Wagner C. J. Biol. Chem. 266:4965-4973(1991).

48. Aldo/keto reductase family signatures

The aldo-keto reductase family [1,2] groups together a number of structurally and functionally related NADPH-dependent oxidoreductases as well as some other proteins. The proteins known to belong to this family are: - Aldehyde reductase (EC 1.1.1.2). - Aldose reductase (EC 1.1.1.21). - 3-alpha-hydroxysteroid dehydrogenase (EC 1.1.1.50), which terminates androgen action by converting 5-alpha-dihydrotestosterone to 3-alpha-androstanediol. - Prostaglandin F synthase (EC 1.1.1.188) which catalyzes the reduction of prostaglandins H2 and D2 to F2-alpha. - D-sorbitol-6-phosphate dehydrogenase (EC 1.1.1.200) from apple. - Morphine 6-dehydrogenase (EC 1.1.1.218) from Pseudomonas putida plasmid pMDH7.2 (gene morA). - Chlordecone reductase (EC 1.1.1.225) which reduces the pesticide chlordecone (kepone) to the corresponding alcohol. - 2,5-diketo-D-gluconic acid reductase (EC 1.1.1.-) which catalyzes the reduction of 2,5-diketogluconic acid to 2-keto-L-gulonic acid, a key intermediate in the production of ascorbic acid. - NAD(P)H-dependent xylose reductase (EC 1.1.1.-) from the yeast Pichia stipitis. This enzyme reduces xylose into xylit. - Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase (EC 1.3.1.20). - 3-oxo-5-beta-steroid 4-dehydrogenase (EC 1.3.99.6) which catalyzes the reduction of delta(4)-3-oxosteroids. - A soybean reductase, which co-acts with chalcone synthase in the formation of 4,2',4'-trihydroxychalcone. - Frog eye lens rho crystallin. - Yeast GCY protein, whose function is not known. - Leishmania major P110/11E protein. P110/11E is a developmentally regulated protein whose abundance is markedly elevated in promastigotes compared with amastigotes. Its exact function is not yet known. - Escherichia coli hypothetical protein yafB. - Escherichia coli hypothetical protein yghE. - Yeast hypothetical protein YBR149w. - Yeast hypothetical protein YHR104w. - Yeast hypothetical protein YJR096w.These proteins have all about 300 amino acid residues. Three consensus patterns have been developed that are specific to this family of proteins. The first one is located in the N-terminal section of these proteins. The second pattern is located in the central section. The third pattern, located in the

C-terminal, is centered on a lysine residue whose chemical modification, in aldose and aldehydereductases, affect the catalytic efficiency.

Consensus pattern: G-[FY]-R-~~[HSAL]~~[HSAL (SEQ ID NO: 269)]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-D-~~[STAGC]~~[STAGC (SEQ ID NO: 691)]-[AS]-x(5)-E-x(2)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]- G -

Consensus pattern: ~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x(9)-~~[KREQ]~~[KREQ (SEQ ID NO: 292)]-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-G-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[SC]-N-[FY]-

Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-~~[PAIV]~~[PAIV (SEQ ID NO: 588)]-[KR]-[ST]-x(4)-R-x(2)-~~[GSTAEQK]~~[GSTAEQK (SEQ ID NO: 225)]-[NSL]-x(2)-~~[LIVMFA]~~[LIVMFA (SEQ ID NO: 403)] [K is a putative active site residue]-

[ 1] Bohren K.M., Bullock B., Wermuth B., Gabbay K.H. J. Biol. Chem. 264:9547-9551(1989).

[ 2] Bruce N.C., Willey D.L., Coulson A.F.W., Jeffery J. Biochem. J. 299:805-811(1994).

49. Alpha amylase. This family is classified as family 13 of the glycosyl hydrolases. The structure is an 8 stranded alpha/beta barrel, interrupted by a ~70 a.a. calcium-binding domain protruding between beta strand 3 and alpha helix 3, and a carboxyl-terminal Greek key beta-barrel domain.

[1] Larson SB, Greenwood A, Cascio D, Day J, McPherson A, J Mol Biol 1994;235:1560-1584.

50. Aminotransferases class-I pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-I, currently consists of the following enzymes: - Aspartate aminotransferase (AAT) (EC 2.6.1.1). AAT catalyzes the reversible transfer of the amino group from L-aspartate to 2-oxoglutarate to form oxaloacetate and L-

glutamate. In eukaryotes, there are two AAT isozymes: one is located in the mitochondrial matrix, the second is cytoplasmic. In prokaryotes, only one form of AAT is found (gene aspC). - Tyrosine aminotransferase (EC 2.6.1.5) which catalyzes the first step in tyrosine catabolism by reversibly transferring its amino group to 2- oxoglutarate to form 4-hydroxyphenylpyruvate and L-glutamate. - Aromatic aminotransferase (EC 2.6.1.57) involved in the synthesis of Phe, Tyr, Asp and Leu (gene tyrB). - 1-aminocyclopropane-1-carboxylate synthase (EC 4.4.1.14) (ACC synthase) from plants. ACC synthase catalyzes the first step in ethylene biosynthesis. - Pseudomonas denitrificans cobC, which is involved in cobalamin biosynthesis. - Yeast hypothetical protein YJL060w.The sequence around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently conserved to allow the creation of a specific pattern.

Consensus pattern: [GS]-~~[LIVMFYTAC]~~[LIVMFYTAC (SEQ ID NO: 462)]-~~[GSTA]~~[GSTA (SEQ ID NO: 217)]-K-x(2)-~~[GSALVN]~~[GSALVN (SEQ ID NO: 203)]-~~[LIVMFA]~~[LIVMFA (SEQ ID NO: 403)]-x-~~[GNAR]~~[GNAR (SEQ ID NO: 184)]- x-R-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)]-[GA] [K is the pyridoxal-P attachment site]

[ 1] Bairoch A. Unpublished observations (1992).

[ 2] Sung M.H., Tanizawa K., Tanaka H., Kuramitsu S., Kagamiyama H., Hirotsu K., Okamoto A., Higuchi T., Soda K. J. Biol. Chem. 266:2567-2572(1991).

51. Aminotransferases class-II pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1] into subfamilies. One of these, called class-II, currently consists of the following enzymes: - Glycine acetyltransferase (EC 2.3.1.29), which catalyzes the addition of acetyl-CoA to glycine to form 2-amino-3-oxobutanoate (gene kbl). - 5-aminolevulinic acid synthase (EC 2.3.1.37) (delta-ALA synthase), which catalyzes the first step in heme biosynthesis via the Shemin (or C4) pathway, i.e. the addition of succinyl-CoA to glycine to form 5-aminolevulinate. - 8-amino-7-oxononanoate synthase (EC 2.3.1.47) (7-KAP synthetase), a bacterial enzyme (gene bioF) which catalyzes an intermediate step in the biosynthesis of biotin: the addition of 6-carboxy-hexanoyl-CoA to alanine to form 8-amino-7-oxononanoate.

- Histidinol-phosphate aminotransferase (EC 2.6.1.9), which catalyzes the eighth step in histidine biosynthetic pathway: the transfer of an amino group from 3-(imidazol-4-yl)-2-oxopropyl phosphate to glutamic acid to form histidinol phosphate and 2-oxoglutarate. - Serine palmitoyltransferase (EC 2.3.1.50) from yeast (genes LCB1 and LCB2), which catalyzes the condensation of palmitoyl-CoA and serine to form 3- ketosphinganine.The sequence around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently conserved to allow the creation of a specific pattern

Consensus pattern: T-~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)]-~~[STAG]~~[STAG (SEQ ID NO: 690)]-K-[SAG]-~~[LIVMFYWR]~~[LIVMFYWR (SEQ ID NO: 479)]-[SAG]-x(2)-[SAG] [K is the pyridoxal-P attachment site]-

[ 1] Bairoch A. Unpublished observations (1991).

52. Aminotransferases class-III pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-III, currently consists of the following enzymes: - Acetylornithine aminotransferase (EC 2.6.1.11) which catalyzes the transfer of an amino group from acetylornithine to alpha-ketoglutarate, yielding N-acetyl-glutamic-5-semi-aldehyde and glutamic acid. - Ornithine aminotransferase (EC 2.6.1.13), which catalyzes the transfer of an amino group from ornithine to alpha-ketoglutarate, yielding glutamic-5- semi-aldehyde and glutamic acid. - Omega-amino acid--pyruvate aminotransferase (EC 2.6.1.18), which catalyzes transamination between a variety of omega-amino acids, mono- and diamines, and pyruvate. It plays a pivotal role in omega amino acids metabolism. - 4-aminobutyrate aminotransferase (EC 2.6.1.19) (GABA transaminase), which catalyzes the transfer of an amino group from GABA to alpha-ketoglutarate, yielding succinate semialdehyde and glutamic acid. - DAPA aminotransferase (EC 2.6.1.62), a bacterial enzyme (gene bioA) which catalyzes an intermediate step in the biosynthesis of biotin, the transamination of 7-keto-8-aminopelargonic acid (7-KAP) to form 7,8- diaminopelargonic acid (DAPA). - 2,2-dialkylglycine decarboxylase (EC 4.1.1.64), a Pseudomonas cepacia enzyme (gene dgdA) that catalyzes the decarboxylating amino transfer of 2,2-dialkylglycine

and pyruvate to dialkyl ketone, alanine and carbon dioxide. - Glutamate-1-semialdehyde aminotransferase (EC 5.4.3.8) (GSA). GSA is the enzyme involved in the second step of porphyrin biosynthesis, via the C5 pathway. It transfers the amino group on carbon 2 of glutamate-1- semialdehyde to the neighbouring carbon, to give delta-aminolevulinic acid. - Bacillus subtilis aminotransferase yhxA. - Bacillus subtilis aminotransferase yodT. - Haemophilus influenzae aminotransferase HI0949. - Caenorhabditis elegans aminotransferase T01B11.2.The sequence around the pyridoxal-phosphate attachment site of this class ofenzyme is sufficiently conserved to allow the creation of a specific pattern.

Consensus pattern: ~~[LIVMFYWC]~~[LIVMFYWC (SEQ ID NO: 465)](2)-x-D-E-[IVA]-x(2)-G-~~[LIVMFAGC]~~[LIVMFAGC (SEQ ID NO: 406)]-x(0,1)- ~~[RSACLI]~~[RSACLI (SEQ ID NO: 648)]-x-~~[GSAD]~~[GSAD (SEQ ID NO: 194)]-x(12,16)-D-~~[LIVMFC]~~[LIVMFC (SEQ ID NO: 412)]-~~[LIVMFYSTA]~~[LIVMFYSTA (SEQ ID NO: 455)]-x(2)- [GSA]-K-x(3)-~~[GSTADNV]~~[GSTADNV (SEQ ID NO: 224)]-~~[GSAC]~~[GSAC (SEQ ID NO: 187)] [K is the pyridoxal-P attachment site]-

[ 1] Bairoch A. Unpublished observations (1992).[ 2] Yonaha K., Nishie M., Aibara S. J. Biol. Chem. 267:12506-12510(1992).

53. Ank repeat. There's no clear separation between noise and signal on the HMM search Ankyrin repeats generally consist of a beta, alpha, alpha, beta order of secondary structures. The repeats associate to form a higher order structure.

[1] A, Holak TA, FEBS Lett 1997;401:127-132.

[2] Lux SE, John KM, Bennett V, Nature 1990;345:736-739.

54. Aminotransferases class-IV signature

Aminotransferases share certain mechanistic features with other pyridoxal-phosphate dependent enzymes, such as the covalent binding of the pyridoxal-phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-IV, currently consists of the following enzymes:

- Branched-chain amino-acid aminotransferase (EC 2.6.1.42) (transaminase B), a bacterial (gene ilvE) and eukaryotic enzyme which catalyzes the reversible transfer of an amino group from 4-methyl-2-oxopentanoate to glutamate, to form leucine and 2-oxoglutarate.
- D-alanine aminotransferase (EC 2.6.1.21). A bacterial enzyme which catalyzes the transfer of the amino group from D-alanine (and other D-amino acids) to 2-oxoglutarate, to form pyruvate and D-aspartate.
- 4-amino-4-deoxychorismate (ADC) lyase (gene pabC). A bacterial enzyme that converts ADC into 4-aminobenzoate (PABA) and pyruvate.

The above enzymes are proteins of about 270 to 415 amino-acid residues that share a few regions of sequence similarity. Surprisingly, the best-conserved region does not include the lysine residue to which the pyridoxal-phosphategroup is known to be attached, in ilvE. The region that has been selected as a signature pattern is located some 40 residues at the C-terminus side of the PlP-lysine

Consensus pattern: E-x-~~[STAGCI]~~[STAGCI (SEQ ID NO: 693)]-x(2)-N-~~[LIVMFAC]~~[LIVMFAC (SEQ ID NO: 404)]-[FY]-x(6,12)-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x-T- x(6,8)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-[GS]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-[KR]-

[1] Green J.M., Merkel W.K., Nichols B.P. J. Bacteriol. 174:5317-5323(1992).
[2] Bairoch A. Unpublished observations (1992).

55. Aminotransferases class-V pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-V, currently consists of the following enzymes: - Phosphoserine aminotransferase (EC 2.6.1.52), an enzyme which catalyzes the reversible interconversion of phosphoserine and 2-oxoglutarate to 3-phosphonooxypyruvate and glutamate. It is required both in the major phosphorylated pathway of serine biosynthesis and in pyridoxine biosynthesis. The bacterial enzyme (gene serC) is highly similar to a rabbit endometrial progesterone-induced protein (EPIP), which is probably a phosphoserine aminotransferase [3]. - Serine--glyoxylate aminotransferase (EC 2.6.1.45) (SGAT) (gene

sgaA) from Methylobacterium extorquens. - Serine--pyruvate aminotransferase (EC 2.6.1.51). This enzyme also acts as an alanine--glyoxylate aminotransferase (EC 2.6.1.44). In vertebrates, it is located in the peroxisomes and/or mitochondria. - Isopenicillin N epimerase (gene cefD). This enzyme is involved in the biosynthesis of cephalosporin antibiotics and catalyzes the reversible isomerization of isopenicillin N and penicillin N. - NifS, a protein of the nitrogen fixation operon of some bacteria and cyanobacteria. The exact function of nifS is not yet known. A highly similar protein has been found in fungi (gene NFS1 or SPL1). - The small subunit of cyanobacterial soluble hydrogenase (EC 1.12.-.-). - Hypothetical protein ycbU from Bacillus subtilis. - Hypothetical protein YFL030w from yeast. The sequence around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently conserved to allow the creation of a specific pattern.

Consensus pattern: ~~[LIVFYCHT]~~[LIVFYCHT (SEQ ID NO: 373)]-[DGH]-~~[LIVMFYAC]~~[LIVMFYAC (SEQ ID NO: 436)]-~~[LIVMFYA]~~[LIVMFYA (SEQ ID NO: 435)]-x(2)-~~[GSTAC]~~[GSTAC (SEQ ID NO: 218)]-~~[GSTA]~~[GSTA (SEQ ID NO: 217)]-[HQR]-K-x(4,6)-G-x-~~[GSAT]~~[GSAT (SEQ ID NO: 207)]-x-~~[LIVMFYSAC]~~[LIVMFYSAC (SEQ ID NO: 453)] [K is the pyridoxal-P attachment site]-

[ 1] Ouzounis C., Sander C. FEBS Lett. 322:159-164(1993).
[ 2] Bairoch A. Unpublished observations (1992).
[ 3] van der Zel A., Lam H.-M., Winkler M.E. Nucleic Acids Res. 17:8379-8379(1989).

56. Annexins repeated domain signature

Annexins [1 to 6] are a group of calcium-binding proteins that associate reversibly with membranes. They bind to phospholipid bilayers in the presence of micromolar free calcium concentration. The binding is specific for calcium and for acidic phospholipids. Annexins have been claimed to be involved in cytoskeletal interactions, phospholipase inhibition, intracellular signalling, anticoagulation, and membrane fusion. Each of these proteins consist of an N-terminal domain of variable length followed by four or eight copies of a conserved segment of sixty one residues. The repeat (sometimes known as an 'endonexin fold') consists of five alpha-helices that are wound into a right-handed superhelix [7].The proteins known to belong to the annexin family are listed below: - Annexin I (Lipocortin 1) (Calpactin 2) (p35) (Chromobindin 9). - Annexin II (Lipocortin 2) (Calpactin 1) (Protein I) (p36) (Chromobindin

8). - Annexin III (Lipocortin 3) (PAP-III). - Annexin IV (Lipocortin 4) (Endonexin I) (Protein II) (Chromobindin 4). - Annexin V (Lipocortin 5) (Endonexin 2) (VAC-alpha) (Anchorin CII) (PAP-I). - Annexin VI (Lipocortin 6) (Protein III) (Chromobindin 20) (p68) (p70). This is the only known annexin that contains 8 (instead of 4) repeats. - Annexin VII (Synexin). - Annexin VIII (Vascular anticoagulant-beta) (VAC-beta). - Annexin IX from Drosophila. - Annexin X from Drosophila. - Annexin XI (Calcyclin-associated annexin) (CAP-50). - Annexin XII from Hydra vulgaris. - Annexin XIII (Intestine-specific annexin) (ISA).The signature pattern for this domain spans positions 9 to 61 of the repeatand includes the only perfectly conserved residue (an arginine in position 22)-

Consensus pattern: [TG]-[STV]-x(8)-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x(2)-R-x(3)-~~[DEQNH]~~[DEQNH (SEQ ID NO: 75)]-x(7)-[IFY]- x(7)-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x(3)-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x(11)-~~[LIVMFA]~~[LIVMFA (SEQ ID NO: 403)]-x(2)-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-

[ 1] Raynal P., Pollard H.B. Biochim. Biophys. Acta 1197:63-93(1994).

[ 2] Barton G.J., Newman R.H., Freemont P.S., Crumpton M.J. Eur. J. Biochem. 198:749-760(1991).

[ 3] Burgoyne R.D., Geisow M.J. Cell Calcium 10:1-10(1989).

[ 4] Haigler H.T., Fitch J.M., Jones J.M., Schlaepfer D.D. Trends Biochem. Sci. 14:48-50(1989).

[ 5] Klee C.B. Biochemistry 27:6645-6653(1988).

[ 6] Smith P.D., Moss S.E. Trends Genet. 10:241-246(1994).

[ 7] Huber R., Roemisch J., Paques E.-P. EMBO J. 9:3867-3874(1990).

[ 8] Fiedler K., Simons K. Trends Biochem. Sci. 20:177-178(1995).

57. (arf_1) ADP-ribosylation factors family signature

ADP-ribosylation factors (ARF) [1,2,3,4] are 20 Kd GTP-binding proteins involved in protein trafficking. They may modulate vesicle budding and uncoating within the Golgi apparatus. ARF's also act as allosteric activators of cholera toxin ADP-ribosyltransferase activity. They are evolutionary conserved and present in all eukaryotes. At least six forms of ARF are present in mammals and three in budding yeast. The ARF family also includes proteins highly related to ARF's but which lack the cholera toxin cofactor activity, they are

collectively known as ARL's (ARF-like).ARD1 is a 64 Kd mammalian protein of unknown biological function that contains an ARF domain at its C-terminal extremity. Proteins from the ARF family are generally included in the RAS 'superfamily' of small GTP-binding proteins [5], but they are only slightly related to the other RAS proteins. They also differ from RAS proteins in that they lack cysteine residues at their C-termini and are therefore not subject to prenylation. The ARFs are N-terminally myristoylated (the ARLs have not yet been shown to be modified in such a fashion). A conserved region in the C-terminal part of ARF's and ARL's has been selected as a signature pattern.

Consensus pattern: ~~[HRQT]~~[HRQT (SEQ ID NO: 267)]-x-~~[FYWI]~~[FYWI (SEQ ID NO: 147)]-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(4)-A-x(2)-G-x(2)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(2)- [GSA]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x-[WK]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-

Note: proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop) (see <PDOC00017

[ 1] Boman A.L., Kahn R.A. Trends Biochem. Sci. 20:147-150(1995).

[ 2] Moss J., Vaughan M. Cell. Signal. 4.367-399(1993).

[ 3] Moss J., Vaughan M. Prog. Nucleic Acid Res. Mol. Biol. 45:47-65(1993).

[ 4] Amor J.C., Harrison D.H., Kahn R.A., Ringe D. Nature 372:704-708(1994).

[ 5] Valencia A., Chardin P., Wittinghofer A., Sander C. Biochemistry 30:4637-4648(1991).

(arf_2) ATP/GTP-binding site motif A (P-loop)

From sequence comparisons and crystallographic data analysis it has been shown [1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix. This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5].There are numerous ATP- or GTP-binding proteins in which the P-loop is found. A number of protein families for which the relevance of the presence of such motif has been noted are listed below: - ATP synthase alpha and beta subunits (see <PDOC00137>). - Myosin heavy chains. - Kinesin heavy chains and kinesin-like proteins (see <PDOC00343>). - Dynamins and dynamin-like proteins (see <PDOC00362>). - Guanylate kinase (see <PDOC00670>). - Thymidine kinase

(see <PDOC00524>). - Thymidylate kinase (see <PDOC01034>). - Shikimate kinase (see <PDOC00868>). - Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>). - ATP-binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>). - DNA and RNA helicases [8,9,10]. - GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.). - Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.). - Nuclear protein ran (see <PDOC00859>). - ADP-ribosylation factors family (see <PDOC00781>). - Bacterial dnaA protein (see <PDOC00771>). - Bacterial recA protein (see <PDOC00131>). - Bacterial recF protein (see <PDOC00539>). - Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.). - DNA mismatch repair proteins mutS family (See <PDOC00388>). - Bacterial type II secretion system protein E (see <PDOC00567>).Not all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.

Consensus pattern: [AG]-x(4)-G-K-[ST]-

[ 1] Walker J.E., Saraste M., Runswick M.J., Gay N.J. EMBO J. 1:945-951(1982).

[ 2] Moller W., Amons R. FEBS Lett. 186:1-7(1985).

[ 3] Fry D.C., Kuby S.A., Mildvan A.S. Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).

[ 4] Dever T.E., Glynias M.J., Merrick W.C. Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).

[ 5] Saraste M., Sibbald P.R., Wittinghofer A. Trends Biochem. Sci. 15:430-434(1990).

[ 6] Koonin E.V. J. Mol. Biol. 229:1165-1174(1993).

[ 7] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J. Bioenerg. Biomembr. 22:571-592(1990).

[ 8] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

[ 9] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).

[10] Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M. Nucleic Acids Res. 17:4713-4730(1989).

58. Arginase family signatures

The following enzymes have been shown [1] to be evolutionary related: - Arginase (EC 3.5.3.1), a ubiquitous enzyme which catalyzes the degradation of arginine to ornithine and urea [2]. - Agmatinase (EC 3.5.3.11) (agmatine ureohydrolase), a prokaryotic enzyme (gene speB) that catalyzes the hydrolysis of agmatine into putrescine and urea. - Formiminoglutamase (EC 3.5.3.8) (formiminoglutamate hydrolase), a prokaryotic enzyme (gene hutG) that hydrolyzes N-formimino-glutamate into glutamate and formamide. - Hypothetical proteins from methanogenic archaebacteria. These enzymes are proteins of about 300 amino-acid residues. Three conserved regions that contain charged residues which are involved in the binding of the two manganese ions [3] can be used as signature patterns.-

Consensus pattern: ~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-G-G-x-H-x-~~[LIVMT]~~[LIVMT (SEQ ID NO: 518)]-~~[STAV]~~[STAV (SEQ ID NO: 733)]-x-[PAG]-x(3)-~~[GSTA]~~[GSTA (SEQ ID NO: 217)] [H binds manganese]-

Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-x-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-D-[AS]-H-x-D [The two D's and the H bind manganese]-

Consensus pattern: [ST]-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-D-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-D-x(3)-[PAQ]-x(3)-P-[GSA]-x(7)-G [The two D's bind manganese]

[ 1] Ouzounis C., Kyrpides N.C. J. Mol. Evol. 39:101-104(1994).

[ 2] Jenkinson C.P., Grody W.W., Cederbaum S.D. Comp. Biochem. Physiol. 114B:107-132(196).

[ 3] Kanyo Z.F., Scolnick L.R., Ash D.E., Christianson D.W. Nature 383:554-557(1996).

59. (asp) Eukaryotic and viral aspartyl proteases active site

Aspartyl proteases, also known as acid proteases, (EC 3.4.23.-) are a widely distributed family of proteolytic enzymes [1,2,3] known to exist invertebrates, fungi, plants, retroviruses and some plant viruses. Aspartate proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue.The two domains most probably evolved from the duplication of an ancestral gene encoding a primordial domain. Currently known eukaryotic aspartyl proteases are: -

Vertebrate gastric pepsins A and C (also known as gastricsin). - Vertebrate chymosin (rennin), involved in digestion and used for making cheese. - Vertebrate lysosomal cathepsins D (EC 3.4.23.5) and E (EC 3.4.23.34). - Mammalian renin (EC 3.4.23.15) whose function is to generate angiotensin I from angiotensinogen in the plasma. - Fungal proteases such as aspergillopepsin A (EC 3.4.23.18), candidapepsin (EC 3.4.23.24), mucoropepsin (EC 3.4.23.23) (mucor rennin), endothiapepsin (EC 3.4.23.22), polyporopepsin (EC 3.4.23.29), and rhizopuspepsin (EC 3.4.23.21). - Yeast saccharopepsin (EC 3.4.23.25) (proteinase A) (gene PEP4). PEP4 is implicated in posttranslational regulation of vacuolar hydrolases. - Yeast barrier pepsin (EC 3.4.23.35) (gene BAR1); a protease that cleaves alpha-factor and thus acts as an antagonist of the mating pheromone. - Fission yeast sxa1 which is involved in degrading or processing the mating pheromones. Most retroviruses and some plant viruses, such as badnaviruses, encode for anaspartyl protease which is an homodimer of a chain of about 95 to 125 amino acids. In most retroviruses, the protease is encoded as a segment of apolyprotein which is cleaved during the maturation process of the virus. It is generally part of the pol polyprotein and, more rarely, of the gagpolyprotein. Conservation of the sequence around the two aspartates of eukaryotic aspartyl proteases and around the single active site of the viral proteases allows us to develop a single signature pattern for both groups of protease.

Consensus pattern: ~~[LIVMFGAC]~~[LIVMFGAC (SEQ ID NO: 416)]-~~[LIVMTADN]~~[LIVMTADN (SEQ ID NO: 520)]-~~[LIVFSA]~~[LIVFSA (SEQ ID NO: 367)]-D-[ST]-G-~~[STAV]~~[STAV (SEQ ID NO: 733)]-~~[STAPDENQ]~~[STAPDENQ (SEQ ID NO: 727)]- x-~~[LIVMFSTNC]~~[LIVMFSTNC (SEQ ID NO: 426)]-x-~~[LIVMFGTA]~~[LIVMFGTA (SEQ ID NO: 418)] [D is the active site residue]

Note: these proteins belong to families A1 and A2 in the classification of peptidases [4,E1

[ 1] Foltmann B. Essays Biochem. 17:52-84(1981).

[ 2] Davies D.R. Annu. Rev. Biophys. Chem. 19:189-215(1990).

[ 3] Rao J.K.M., Erickson J.W., Wlodawer A. Biochemistry 30:4663-4671(1991).

[ 4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:105-120(1995).

60. (BIRA) Biotin repressor

[1] Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ, Matthews BW; Proc Natl Acad Sci USA 1992;89:9257-9261.

61. BTB/POZ domain

The BTB (for BR-C, ttk and bab) [1] or POZ (for Pox virus and Zinc finger)[2] domain is present near the N-terminus of a fraction of zinc finger (zf-C2H2) proteins and in proteins that contain the Kelch motif such as Kelch and a family of pox virus proteins. The BTB/POZ domain mediates homomeric dimerisation and in some instances heteromeric dimerisation [2].The structure of the dimerised PLZF BTB/POZ domain has been solved and consists of a tightly intertwined homodimer. The central scaffolding of the protein is made up of a cluster of alpha-helices flanked by short beta-sheets at both the top and bottom of the molecule [3]. POZ domains from several zinc finger proteins have been shown to mediate transcriptional repression and to interact with components of histone deacetylase co-repressor complexes including N-CoR and SMRT [4,5,6]. The POZ or BTB domain is also known as BR-C/Ttk or ZiN

[1] Zollman S, Godt D, Prive GG, Couderc JL, Laski FA; Proc Natl Acad Sci U S A 1994;91:10717-10721.

[2]Bardwell VJ, Treisman R; Genes Dev 1994;8:1664-1677.

[3] Ahmad KF, Engel CK, Prive GG; Proc Natl Acad Sci U S A 1998;95:12123-12128.

[4] Deweindt C, Albagli O, Bernardin F, Dhordain P, Quief S, Lantoine D, Kerckaert JP, Leprince D; Cell Growth Differ 1995;6:1495-1503.

[5] Huynh KD, Bardwell VJ; Oncogene 1998;17:2473-2484.

[6] Wong CW, Privalsky ML; J Biol Chem 1998;273:27695-27702.

62. (Bac GSPproteins) Bacterial type II secretion system protein D signature

A number of bacterial proteins, some of which are involved in a general secretion pathway (GSP) for the export of proteins (also called the type II pathway) [1 to 5], have been found to be evolutionary related. These proteins are listed below: - The 'D' protein from the GSP operon of: Aeromonas (gene exeD); Erwinia (gene outD); Escherichia coli (gene yheF), Klebsiella pneumoniae (gene pulD); Pseudomonas aeruginosa (gene xcpQ); Vibrio cholerae (gene epsD) and Xanthomonas campestris (gene xpsD). - comE from Haemophilus influenzae, involved in competence (DNA uptake). - pilQ from Pseudomonas aeruginosa, which is essential for the formation of the pili. - hofQ (hopQ) from Escherichia coli. - hrpH

from Pseudomonas syringae, which is involved in the secretion of a proteinaceous elicitor of the hypersensitivity response in plants. - hrpA1 from Xanthomonas campestris pv. vesicatoria, which is also involved in the hypersensitivity response. - mxiD from Shigella flexneri which is involved in the secretion of the Ipa invasins which are necessary for penetration of intestinal epithelial cells. - omc from Neisseria gonorrhoeae. - yssC from Yersinia enterocolitica virulence plasmid pYV, which seems to be required for the export of the Yop virulence proteins. - The gpIV protein from filamentous phages such as f1, ike, or m13. GpIV is said to be involved in phage assembly and morphogenesis. These proteins all seem to start with a signal sequence and are thought to be integral proteins in the outer membrane. As a signature pattern a conserved region in the C-terminal section of these proteins has been selected

Consensus pattern: [GR]-~~[DEQKG]~~[DEQKG (SEQ ID NO: 72)]-~~[STVM]~~[STVM (SEQ ID NO: 760)]-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)](3)-[GA]-G-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x(11)- ~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-P-~~[LIVMFYWGS]~~[LIVMFYWGS (SEQ ID NO: 473)]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-~~[GSAE]~~[GSAE (SEQ ID NO: 197)]-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-P- ~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)](2)-x(2)-[LV]-F

[ 1] Salmond G.P.C., Reeves P.J. Trends Biochem. Sci. 18:7-12(1993).

[ 2] Reeves P.J., Whitcombe D., Wharam S., Gibson M., Allison G., Bunce N., Barallon R., Douglas P., Mulholland V., Stevens S., Walker S., Salmond G.P.C. Mol. Microbiol. 8:443-456(1993).

[ 3] Martin P.R., Hobbs M., Free P.D., Jeske Y., Mattick J.S. Mol. Microbiol. 9:857-868(1993).

[ 4] Hobbs M., Mattick J.S. Mol. Microbiol. 10:233-243(1993).

[ 5] Genin S., Boucher C.A. Mol. Gen. Genet. 243:112-118(1994).

63. (Bac globin) Protozoan/cyanobacterial globins signature

Globins are heme-containing proteins involved in binding and/or transporting oxygen [1]. Almost all globins belong to a large family (see <PDOC00793>), the only exceptions are the following proteins which form a family of their own[2,3]: - Monomeric hemoglobins from the protozoan Paramecium caudatum, Tetrahymena pyriformis and Tetrahymena

thermophila. - Cyanoglobin from the cyanobacteria Nostoc commune. - Globins LI637 and LI410 from the chloroplast of the alga Chlamydomonas eugametos. - Mycobacterium tuberculosis hypothetical protein MtCY48.23.These proteins contain a conserved histidine which could be involved in heme-binding. As a signature pattern, a conserved region that ends with this residue was used

Consensus pattern: F-[LF]-x(5)-G-[PA]-x(4)-G-[KRA]-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(3)-H-

[ 1] Concise Encyclopedia Biochemistry, Second Edition, Walter de Gruyter, Berlin New-York (1988).

[ 2] Takagi T. Curr. Opin. Struct. Biol. 3:413-418(1993).

[ 3] Couture M., Chamberland H., St-Pierre B., Lafontaine J., Guertin M.; Mol. Gen. Genet. 243:185-197(1994).

64. Band 7 protein family signature

Mammalian band 7 protein [1] (also known as 7.2B or stomatin) is an integral membrane phosphoprotein of red blood cells thought to regulate cation conductance by interacting with other proteins of the junctional complex of the membrane skeleton. Structurally, band 7 is evolutionary related to the following proteins: - Caenorhabditis elegans protein mec-2 [2]. Mec-2 positively regulates the activity of the putative mechanosensory transduction channel. It may links the mechanosensory channel and the microtubule cytoskeleton of the touch receptor neurons. - Caenorhabditis elegans proteins sto-1 to sto-4. - Caenorhabditis elegans protein unc-1. - Escherichia coli hypothetical protein ybbK. - Mycobacterium tuberculosis hypothetical protein MtCY277.09. - Synechocystis strain PCC 6803 hypothetical protein slr1128. - Methanococcus jannaschii hypothetical protein MJ0827.Structurally all these proteins consist of a short N-terminal domain which is followed by a transmembrane region and a variable size (from 170 to 350residues) C-terminal domain .As a signature pattern, a conserved region located about 110residues after the transmembrane domain was selected

Consensus pattern: R-x(2)-[LIV]-[SAN]-x(6)-[LIV]-D-x(2)-T-x(2)-W-G-[LIV]- [KRH]-[LIV]-x-[KR]-[LIV]-E-[LIV]-[KR]-

[ 1] Gallagher P.G., Forget B.G. J. Biol. Chem. 270:26358-26363(1995).

[ 2] Huang M., Gu G., Ferguson E.L., Chalfie M. Nature 378:292-295(1995).

65. Barwin domain signatures

Barwin [1] is a barley seed protein of 125 residues that binds weakly a chitinanalog. It contains six cysteines involved in disulfide bonds, as shown in the following schematic representation.

+----------+ | ***** | ****

xxxxxxxxxxxxxxxCxxxxxxxxxxCxxxxCxCxxxxxxxxCxxxxxxxxxxxxxxxxxxxCx | | | | +----------+ +-------------------------+'C': conserved cysteine involved in a disulfide bond.'*': position of the patterns. Barwin is closely related to the following proteins: - Hevein, a wound-induced protein found in the latex of rubber trees. - HEL, an Arabidopsis thaliana hevein-like protein [2]. - Win1 and win2, two wound-induced proteins from potato. - Pathogenesis-related protein 4 from tobacco. Hevein and the win1/2 proteins consist of an N-terminal chitin-binding domain followed by a barwin-like C-terminal domain. Barwin and its related proteins could be involved in a defense mechanism in plants. As signature patterns, two highly conserved regions that contain some of the cysteines were selected

Consensus pattern: C-G-[KR]-C-L-x-V-x-N [The two C's are involved in disulfide bonds]-
Consensus pattern: V-[DN]-Y-[EQ]-F-V-[DN]-C [C is involved in a disulfide bond]-

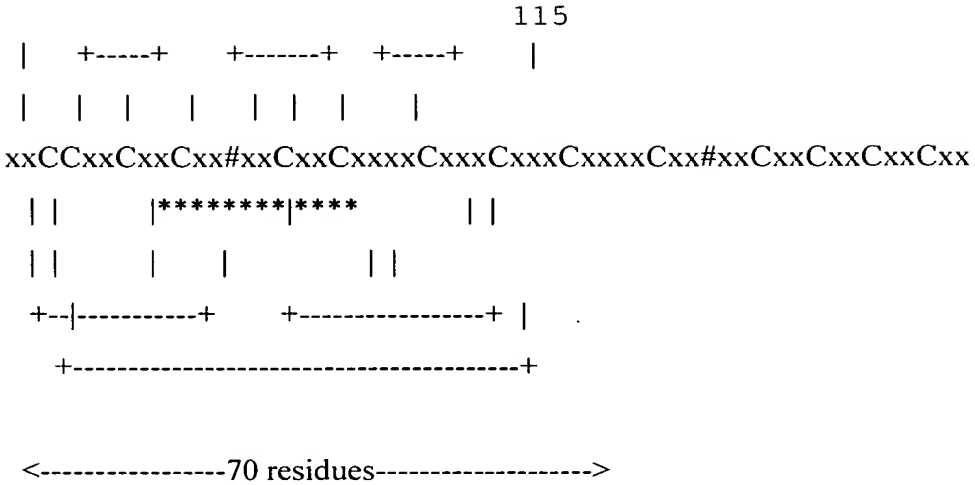
[ 1] Svensson B., Svendsen I., Hoejrup P., Roepstorff P., Ludvigsen S., Poulsen F.M. Biochemistry 31:8767-8770(1992).

[ 2] Potter S., Uknes S., Lawton K., Winter A.M., Chandler D., Dimaio J., Novitzky R., Ward E., Ryals J. Mol. Plant Microbe Interact. 6:680-685(1993).

66. (Bowman-Birk leg) Bowman-Birk serine protease inhibitors family signature

PROSITE cross-reference(s). The Bowman-Birk inhibitor family [1] is one of the numerous families of serine proteinase inhibitors. As it can be seen in the schematic representation, they have a duplicated structure and generally possess two distinct inhibitory sites:

+----------+




These inhibitors are found in the seeds of all leguminous plants as well as in cereal grains. In cereals they exist in two forms, one of which is a duplication of the basic structure shown above [2]. The pattern that was developed to pick up sequences belonging to this family of inhibitors is in the central part of the domain and includes four cysteines.

Consensus pattern C-x(5,6)-~~[DENQKRHSTA]~~[DENQKRHSTA (SEQ ID NO: 42)]-C-~~[PASTDH]~~[PASTDH (SEQ ID NO: 593)]-~~[PASTDK]~~[PASTDK (SEQ ID NO: 594)]-~~[ASTDV]~~[ASTDV (SEQ ID NO: 11)]-C-~~[NDKS]~~[NDKS (SEQ ID NO: 562)]-~~[DEKRHSTA]~~[DEKRHSTA (SEQ ID NO: 24)]-C [The four C's are involved in disulfide bonds] Note this pattern can be found twice in some duplicated cereal inhibitors.

[ 1] Laskowski M., Kato I. Annu. Rev. Biochem. 49:593-626(1980).
[ 2] Tashiro M., Hashino K., Shiozaki M., Ibuki F., Maki Z. J. Biochem. 102:297-306(1987).

67. Pathogenesis-related protein Bet v I family signature

A number of plant proteins, which all seem to be involved in pathogen defense response, are structurally related [1,2,3]. These proteins are:

- Bet v I, the major pollen allergen from white birch. Bet v I is the main cause of type I allergic reactions in Europe, North America and USSR.
- Aln g I, the major pollen allergen from alder.
- Api G I, the major allergen from celery.
- Car b I, the major pollen allergen from hornbeam.
- Cor a I, the major pollen allergen from hazel.
- Mal d I, the major pollen allergen from apple.
- Asparagus wound-induced protein AoPR1.
- Kidney bean pathogenesis-related proteins 1 and 2.
- Parsley pathogenesis-related proteins PR1-1 and PR1-3.
- Pea disease resistance response proteins pI49, pI176 and DRRG49-C.
- Pea abscisic acid-responsive proteins ABR17 and ABR18.
- Potato pathogenesis-related proteins STH-2 and STH-21.
- Soybean stress-induced protein SAM22.

These proteins are thought to be intracellularly located. They contain from 155 to 160 amino acid residues. As a signature pattern, a conserved region located in the third quarter of these proteins has been selected

Consensus pattern: G-x(2)-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x(4)-E-x(2)-~~[CSTAEN]~~[CSTAEN (SEQ ID NO: 18)]-x(8,9)-[GND]-G-[GS]- [CS]-x(2)-K-x(4)-[FY]-

[1] Breiteneder H., Pettenburger K., Bito A., Valenta R., Kraft D., Rumpold H., Scheiner O., Breitenbach M. EMBO J. 8:1935-1938(1989).

[2] Crowell D., John M.E., Russell D., Amasino R.M. Plant Mol. Biol. 18:459-466(1992).

[3] Warner S.A.J., Scott R., Draper J. Plant Mol. Biol. 19:555-561(1992).

68. bZIP transcription factors basic domain signature

The bZIP superfamily [1,2,] of eukaryotic DNA-binding transcription factors groups together proteins that contain a basic region mediating sequence-specific DNA-binding followed by a leucine zipper required for dimerization. This family is quite large, therefore only a parital list of some representative members appears here. - Transcription factor AP-1, which binds selectively to enhancer elements in the cis control regions of SV40 and metallothionein IIA. AP-1, also known as c-jun, is the cellular homolog of the avian sarcoma virus 17 (ASV17) oncogene v-jun. - Jun-B and jun-D, probable transcription factors which are highly similar to

jun/AP-1. - The fos protein, a proto-oncogene that forms a non-covalent dimer with c-jun. - The fos-related proteins fra-1, and fos B. - Mammalian cAMP response element (CRE) binding proteins CREB, CREM, ATF-1, ATF-3, ATF-4, ATF-5, ATF-6 and LRF-1. - Maize Opaque 2, a trans-acting transcriptional activator involved in the regulation of the production of zein proteins during endosperm. - Arabidopsis G-box binding factors GBF1 to GBF4, Parsley CPRF-1 to CPRF-3, Tobacco TAF-1 and wheat EMBP-1. All these proteins bind the G-box promoter elements of many plant genes. - Drosophila protein Giant, which represses the expression of both the kruppel and knirps segmentation gap genes. - Drosophila Box B binding factor 2 (BBF-2), a transcriptional activator that binds to fat body-specific enhancers of alcohol dehydrogenase and yolk protein genes. - Drosophila segmentation protein cap'n'collar (gene cnc), which is involved in head morphogenesis. - Caenorhabditis elegans skn-1, a developmental protein involved in the fate of ventral blastomeres in the early embryo. - Yeast GCN4 transcription factor, a component of the general control system that regulates the expression of amino acid-synthesizing enzymes in response to amino acid starvation, and the related Neurospora crassa cpc-1 protein. - Neurospora crassa cys-3 which turns on the expression of structural genes which encode sulfur-catabolic enzymes. - Yeast MET28, a transcriptional activator of sulfur amino acids metabolism. - Yeast PDR4 (or YAP1), a transcriptional activator of the genes for some oxygen detoxification enzymes. - Epstein-Barr virus trans-activator protein BZLF1.-

Consensus pattern: [KR]-x(1,3)-~~[RKSAQ]~~[RKSAQ (SEQ ID NO: 646)]-N-x(2)-[SAQ](2)-x-~~[RKTAENQ]~~[RKTAENQ (SEQ ID NO: 647)]-x-R-x-[RK]-

[ 1] Hurst H.C. Protein Prof. 2:105-168(1995).[ 2] Ellenberger T. Curr. Opin. Struct. Biol. 4:12-21(1994).

69. Biotin-requiring enzymes attachment site

Biotin, which plays a catalytic role in some carboxyl transfer reactions, is covalently attached, via an amide bond, to a lysine residue in enzymes requiring this coenzyme [1,2,3,4]. Such enzymes are:

- Pyruvate carboxylase (EC 6.4.1.1).
- Acetyl-CoA carboxylase (EC 6.4.1.2).
- Propionyl-CoA carboxylase (EC 6.4.1.3).

- Methylcrotonoyl-CoA carboxylase (EC 6.4.1.4).
- Geranoyl-CoA carboxylase (EC 6.4.1.5).
- Urea carboxylase (EC 6.3.4.6).
- Oxaloacetate decarboxylase (EC 4.1.1.3).
- Methylmalonyl-CoA decarboxylase (EC 4.1.1.41).
- Glutaconyl-CoA decarboxylase (EC 4.1.1.70).
- Methylmalonyl-CoA carboxyl-transferase (EC 2.1.3.1) (transcarboxylase).

Sequence data reveal that the region around the biocytin (biotin-lysine) residue is well conserved and can be used as a signature pattern.

Consensus pattern[GN]-~~[DEQTR]~~[DEQTR (SEQ ID NO: 81)]-x-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x(2)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-[AIV]-M-K-~~[LMAT]~~[LMAT (SEQ ID NO: 541)]-x(3)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-[SAV] [K is the biotin attachment site] Note the domain around the biotin-binding lysine residue is evolutionary related to that around the lipoyl-binding lysine residue of 2-oxo acid dehydrogenase acyltransferases

[ 1] Knowles J.R. Annu. Rev. Biochem. 58:195-221(1989).
[ 2] Samols D., Thronton C.G., Murtif V.L., Kumar G.K., Haase F.C., Wood H.G. J. Biol. Chem. 263:6461-6464(1988).
[ 3] Goss N.H., Wood H.G. Meth. Enzymol. 107:261-278(1984).
[ 4] Shenoy B.C., Xie Y., Park V.L., Kumar G.K., Beegen H., Wood H.G., Samols D. J. Biol. Chem. 267:18407-18412(1992).

2-oxo acid dehydrogenases acyltransferase component lipoyl binding site

The 2-oxo acid dehydrogenase multienzyme complexes [1,2] from bacterial and eukaryotic sources catalyze the oxidative decarboxylation of 2-oxo acids to the corresponding acyl-CoA. The three members of this family of multienzyme complexes are:

- Pyruvate dehydrogenase complex (PDC).
- 2-oxoglutarate dehydrogenase complex (OGDC).
- Branched-chain 2-oxo acid dehydrogenase complex (BCOADC).

These three complexes share a common architecture: they are composed of multiple copies of three component enzymes - E1, E2 and E3. E1 is a thiamine pyrophosphate-dependent 2-oxo acid dehydrogenase, E2 a dihydrolipamide

acyltransferase, and E3 an FAD-containing dihydrolipamide dehydrogenase. E2 acyltransferases have an essential cofactor, lipoic acid, which is covalently bound via a amide linkage to a lysine group. The E2 components of OGCD and BCOACD bind a single lipoyl group, while those of PDC bind either one (in yeast and in Bacillus), two (in mammals), or three (in Azotobacter and in Escherichia coli) lipoyl groups [3].

In addition to the E2 components of the three enzymatic complexes described above, a lipoic acid cofactor is also found in the following proteins:

- H-protein of the glycine cleavage system (GCS) [4]. GCS is a multienzyme complex of four protein components, which catalyzes the degradation of glycine. H protein shuttles the methylamine group of glycine from the P protein to the T protein. H-protein from either prokaryotes or eukaryotes binds a single lipoic group.
- Mammalian and yeast pyruvate dehydrogenase complexes differ from that of other sources, in that they contain, in small amounts, a protein of unknown function - designated protein X or component X. Its sequence is closely related to that of E2 subunits and seems to bind a lipoic group [5].
- Fast migrating protein (FMP) (gene acoC) from Alcaligenes eutrophus [6]. This protein is most probably a dihydrolipamide acyltransferase involved in acetoin metabolism.

A signature pattern was developed which allows the detection of the lipoyl-binding site.

Consensus pattern[GN]-x(2)-~~[LIVF]~~[LIVF (SEQ ID NO: 360)]-x(5)-~~[LIVFC]~~[LIVFC (SEQ ID NO: 362)]-x(2)-~~[LIVFA]~~[LIVFA (SEQ ID NO: 361)]-x(3)-K-~~[STAIV]~~[STAIV (SEQ ID NO: 712)]-~~[STAVQDN]~~[STAVQDN (SEQ ID NO: 736)]-x(2)-~~[LIVMFS]~~[LIVMFS (SEQ ID NO: 422)]-x(5)-[GCN]-x-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)] [K is the lipoyl-binding site] Note the domain around the lipoyl-binding lysine residue is evolutionary related to that around the biotin-binding lysine residue of biotin requiring enzymes

[1] Yeaman S.J. Biochem. J. 257:625-632(1989).

[2] Yeaman S.J. Trends Biochem. Sci. 11:293-296(1986).

[3] Russel G.C., Guest J.R. Biochim. Biophys. Acta 1076:225-232(1991).

[4] Fujiwara K., Okamura-Ikeda K., Motokawa Y. J. Biol. Chem. 261:8836-8841(1986).

[ 5] Behal R.H., Browning K.S., Hall T.B., Reed L.J. Proc. Natl. Acad. Sci. U.S.A. 86:8732-8736(1989).

[ 6] Priefert H., Hein S., Krueger N., Zeh K., Schmidt B., Steinbuechel A. J. Bacteriol. 173:4056-4071(1991).

70. C2 (C2 domain) Number of members: 295

Some isozymes of protein kinase C (PKC) [1,2] contain a domain, known as C2, of about 116 amino-acid residues which is located between the two copies of the C1 domain (that bind phorbol esters and diacylglycerol) (see <PDOC00379>) and the protein kinase catalytic domain (see <PDOC00100>). Regions with significant homology [3,E1] to the C2-domain have been found in the following proteins:

- PKC isoforms alpha, beta and gamma and Drosophila isoforms PKC1 and PKC2.
- PKC isoforms delta, epsilon and eta, Caenorhabditis elegans kin-13 and yeast PKC1 have a C2-like domain at the N-terminal extremity [4].
- Yeast cAMP dependent protein kinase SCH9 contains a C2-like domain.
- Mammalian phosphatidylinositol-specific phospholipase C (PI-PLC) (see <PDOC50007>) isoforms beta, gamma and delta as well as several non-mammalian PI-PLCs have a C2-like domain C-terminal of the catalytic domain.
- Mammalian and plants phosphatidylinositol-3-kinase have a C2-like domain in the central region of the 110 Kd catalytic subunit.
- Yeast phosphatidylserine-decarboxylase 2 (gene PSD2) contains a C2 domain in its central region.
- Cytosolic phospholipase D from plants and cytosolic phospholipase A2 have a C2-like domain at their N-terminus.
- Synaptotagmins (p65). This is a family of related synaptic vesicle proteins that bind acidic phospholipids and that may have a regulatory role in the membrane interactions during trafficking of synaptic vesicles at the active zone of the synapse. All isoforms of synaptotagmins have two copies of the C2 domain in their C-terminal region.
- Rabphilin-3A, a synaptic protein contains two C2 domains.
- Caenorhabditis elegans protein unc-13 whose function is not known. Unc-13 has a C2 domain in its central part and a C2-like domain at the C-terminus.
- rasGAP and the breakpoint cluster protein bcr have a C2-domain C-terminal of a PH-domain.

- Yeast protein BUD2 (or CLA2) has a C2-domain in the central region.
- Yeast protein RSP5 and human protein NEDD-4, both proteins also contain WW domains (see <PDOC50020>).
- Perforin (see <PDOC00251>) has a C2 domain at the C-terminus. It is the only extracellular protein known to contain a C2 domain.
- Yeast hypothetical protein YML072C has a C2 domain.
- Yeast hypothetical protein YNL087W has three C2 domains.
- Caenorhabditis elegans hypothetical protein F37A4.7 has two C2 domains.

The C2 domain is thought to be involved in calcium-dependent phospholipid binding [5]. Since domains related to the C2 domain are also found in proteins that do not bind calcium, other putative functions for the C2 domain like e.g. binding to inositol-1,3,4,5-tetraphosphate have been suggested [6]. Recently, the 3D structure of the first C2 domain of synaptotagmin has been reported [7], the domain forms an eight-stranded beta sandwich. The signature pattern that has been developed for the C2 domain is located in a conserved part of that domain, the connecting loop between beta strands 2 and 3. A profile has been developed for the C2 domain that covers the total domain.

-Consensus pattern: [ACG]-x(2)-L-x(2,3)-D-x(1,2)-~~[NGSTLIF]~~[NGSTLIF (SEQ ID NO: 567)]-~~[GTMR]~~[GTMR (SEQ ID NO: 256)]-x-~~[STAP]~~[STAP (SEQ ID NO: 726)]-D-[PA]-[FY]

-Note: this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.

[1]Medline: 96367095 Extending the C2 domain family: C2s in PKCs delta, epsilon, eta and theta, phospholipases, GAPs and perforin. Ponting CP, Parker PJ; Protein Sci 1996;5:162-166.

[ 1] Azzi A., Boscoboinik D., Hensey C. Eur. J. Biochem. 208:547-557(1992).

[ 2] Stabel S. Semin. Cancer Biol. 5:277-284(1994).

[ 3] Brose N., Hofmann K.O., Hata Y., Suedhof T.C. J. Biol. Chem. 270:25273-25280(1995).

[ 4] Sossin W.S., Schwartz J.H. Trends Biochem. Sci. 18:207-208(1993).

[ 5] Davletov B.A., Suedhof T.C. J. Biol. Chem. 268:26386-26390(1993).

[ 6] Fukuda M., Aruga J., Niinobe M., Aimoto S., Mikoshiba K. J. Biol. Chem. 269:29206-29211(1994).

[ 6] Sutton R.B., Davletov B.A., Berghuis A.M., Suedhof T.C., Sprang S.R. Cell 80:929-938(1995).

71. CAP (CAP protein) Number of members: 11

In budding and fission yeasts the CAP protein is a bifunctional protein whose N-terminal domain binds to adenylyl cyclase, thereby enabling that enzyme to be activated by upstream regulatory signals, such as Ras. The function of the C-terminal domain is less clear, but it is required for normal cellular morphology and growth control [1]. CAP is conserved in higher eukaryotic organisms where its function is not yet clear [2].

Structurally, CAP is a protein of 474 to 551 residues which consist of two domains separated by a proline-rich hinge. Two signature patterns, one corresponding to a conserved region in the N-terminal extremity and the other to a C-terminal region have been developed.

-Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-x-R-L-[DE]-x(4)-R-L-E

-Consensus pattern: D-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x-E-x-[PA]-x-P-E-Q-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-K

[ 1] Kawamukai M., Gerst J., Field J., Riggs M., Rodgers L., Wigler M., Young D. Mol. Biol. Cell 3:167-180(1992).

[ 2] Yu G., Swiston J., Young D. J. Cell Sci. 107:1671-1678(1994).

72. CAP_GLY (CAP-Gly domain)

CAP stands for cytoskeleton-associated proteins. Swiss:P39937 may be a member but has not been included. It has a weak match to the family between residues 22-67. Number of members: 24

[1]Medline: 93242656. Sequence homologies between four cytoskeleton-associated proteins. Riehemann K, Sorg C; Trends Biochem Sci 1993;18:82-83.

It has been shown [1] that some cytoskeleton-associated proteins (CAP) share the presence of a conserved, glycine-rich domain of about 42 residues, called here CAP-Gly. Proteins known to contain this domain are listed below.

- Restin (also known as cytoplasmic linker protein-170 or CLIP-170), a 160 Kd protein associated with intermediate filaments and that links endocytic vesicles to microtubules. Restin contains two copies of the CAP-Gly domain.

- Vertebrate dynactin (150 Kd dynein-associated polypeptide; DAP) and Drosophila glued, a major component of activator I, a 20S polypeptide complex that stimulates dynein-mediated vesicle transport.

- Yeast protein BIK1 which seems to be required for the formation or stabilization of microtubules during mitosis and for spindle pole body fusion during conjugation.

- Yeast protein NIP100 (NIP80).

- Human protein CKAP1/TFCB, Schizosaccharomyces pombe protein alp11 and Caenorhabditis elegans hypothetical protein F53F4.3. These proteins contain a N-terminal ubiquitin domain (see <PDOC00271>) and a C-terminal CAP-Gly domain.

- Caenorhabditis elegans hypothetical protein M01A8.2.

- Yeast hypothetical protein YNL148c.

Structurally, these proteins are made of three distinct parts: an N-terminal section that is most probably globular and contains the CAP-Gly domain, a large central region predicted to be in an alpha-helical coiled-coil conformation and, finally, a short C-terminal globular domain. The signature for the CAP-Gly domain corresponds to the first 32 residues of the domain and includes five of the six conserved glycines.

-Consensus pattern: G-x(8,10)-[FYW]-x-G-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x(4)-G-K-[NH]-x-G-~~[STAR]~~[STAR (SEQ ID NO: 732)]-x(2)-G-x(2)-[LY]-F

[1] Riehemann K., Sorg C. Trends Biochem. Sci. 18:82-83(1993).

73. (CBD 1)

Cellulose-binding domain, fungal type

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1].

Structurally, cellulases and xylanases generally consist of a catalytic domain joined to a cellulose-binding domain (CBD) by a short linker sequence rich in proline and/or hydroxy-amino acids.

The CBD of a number of fungal cellulases has been shown to consist of 36 amino acid residues. Enzymes known to contain such a domain are:

- Endoglucanase I (gene egl1) from Trichoderma reesei.
- Endoglucanase II (gene egl2) from Trichoderma reesei.
- Endoglucanase V (gene egl5) from Trichoderma reesei.
- Exocellobiohydrolase I (gene CBHI) from Humicola grisea, Neurospora crassa, Phanerochaete chrysosporium, Trichoderma reesei, and Trichoderma viride.
- Exocellobiohydrolase II (gene CBHII) from Trichoderma reesei.
- Exocellobiohydrolase 3 (gene cel3) from Agaricus bisporus
- Endoglucanases B, C2, F and K from Fusarium oxysporum.

The CBD domain is found either at the N-terminal (Cbh-II or egl2) or at the C-terminal extremity (Cbh-I, egl1 or egl5) of these enzymes. As it is shown in the following schematic representation, there are four conserved cysteines in this type of CBD domain, all involved in disulfide bonds.

```
          +----------------+
          |       +------|---------+
          |       |  |       |
xxxxxxxCxxxxxxxxxxCxxxxxCxxxxxxxxxCx
          ****************************
```

'C': conserved cysteine involved in a disulfide bond.

'*': position of the pattern.

Such a domain has also been found in a putative polysaccharide binding protein from the red alga, Porphyra purpurea [2]. Structurally, this protein consists of four tandem repeats of the CBD domain.

Consensus patternC-G-G-x(4,7)-G-x(3)-C-x(5)-C-x(3,5)-[NHG]-x-~~[FYWM]~~[FYWM (SEQ ID NO: 155)]- x(2)-Q-C [The four C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern ALL.

[ 1] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[ 2] Liu Q., der Meer J.P., Reith M.E.

74. CBS domain. 3D Structure found as a subdomain in TIM barrel of inosine-. CBS domain web page. CBS domains are small intracellular modules mostly found in 2 or four copies within a protein. CBS domains are found in cystathionine-beta-synthase (CBS) where mutations lead to homocystinuria. Two CBS domains are found in inosine-monophosphate dehydrogenase from all species, however the CBS domains are not needed for activity. Two CBS domains are found in intracellular loops of several chloride channels. Mutations in this domain of Swiss:P35520 lead to homocystinuria.

Number of members: 414

[1]Medline: 97172695 The structure of a domain common to archaebacteria and the homocystinuria disease protein. Bateman A; Trends Biochem Sci 1997;22:12-13.

[2]Medline: 96279836 Structure and mechanism of inosine monophosphate dehydrogenase in complex with the immunosuppressant mycophenolic-acid. Sintchak MD, Fleming MA, Futer O, Raybuck SA, Chambers SP, Caron PR, Murcko MA, Wilson KP; Cell 1996;85:921-930.

Discovery of CBS domain.

[3]Medline: 97259972 CBS domains in ClC chloride channels implicated in myotonia and nephrolithiasis (kidney stones). Ponting CP; J Mol Med 1997;75:160-163.

75. CDP-OH_P_transf (CDP-alcohol phosphatidyltransferase)

All of these members have the ability to catalyze the displacement of CMP from a CDP-alcohol by a second alcohol with formation of a phosphodiester bond and concomitant breaking of a phosphoride anhydride bond. Number of members: 32

A number of phosphatidyltransferases, which are all involved in phospholipid biosynthesis and that share the property of catalyzing the displacement of CMP from a CDP-alcohol by a second alcohol with formation of a phosphodiester bond and concomitant breaking of a phosphoride anhydride bond share a conserved sequence region [1,2]. These enzymes are:

- Ethanolaminephosphotransferase (EC 2.7.8.1) from yeast (gene EPT1).
- Diacylglycerol cholinephosphotransferase (EC 2.7.8.2) from yeast (gene CPT1).
- Phosphatidylglycerophosphate synthase (EC 2.7.8.5) (CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase) from bacteria (gene pgsA).
- Phosphatidylserine synthase (EC 2.7.8.8) (CDP-diacylglycerol--serine O-phosphatidyltransferase) from yeast (gene CHO1) and from Bacillus subtilis (gene pssA).
- Phosphatidylinositol synthase (EC 2.7.8.11) (CDP-diacylglycerol--inositol 3-phosphatidyltransferase) from yeast (gene PIS).

These enzymes are proteins of from 200 to 400 amino acid residues. The conserved region contains three aspartic acid residues and is located in the N-terminal section of the sequences.

-Consensus pattern: D-G-x(2)-A-R-x(8)-G-x(3)-D-x(3)-D

[1]Medline: 97075020 Two-dimensional 1H-NMR of transmembrane peptides from Escherichia coli phosphatidylglycerophosphate synthase in micelles. Morein S, Trouard TP, Hauksson JB, Rilfors L, Arvidson G, Lindblom G; Eur J Biochem 1996;241:489-497.

[ 1] Nikawa J.-I., Kodaki T., Yamashita S.
J. Biol. Chem. 262:4876-4881(1987).

[ 2] Hjelmstad R.H., Bell R.M.
J. Biol. Chem. 266:5094-5134(1991).

76. CHOD (Cholesterol oxidase) Members of the GMC oxidoreductase family. Number of members: 3

[1]Medline: 94032271. Crystal structure of cholesterol oxidase complexed with a steroid substrate: implications for flavin adenine dinucleotide dependent alcohol oxidases. Li J, Vrielink A, Brick P, Blow DM; Biochemistry 1993;32:11507-11515.

The following FAD flavoproteins oxidoreductases have been found [1,2] to be evolutionary related. These enzymes, which are called 'GMC oxidoreductases', are listed below.

- Glucose oxidase (EC 1.1.3.4) (GOX) from Aspergillus niger. Reaction catalyzed: glucose + oxygen -> delta-luconolactone + hydrogen peroxide.
- Methanol oxidase (EC 1.1.3.13) (MOX) from fungi. Reaction catalyzed: methanol + oxygen -> acetaldehyde + hydrogen peroxide.
- Choline dehydrogenase (EC 1.1.99.1) (CHD) from bacteria. Reaction catalyzed: choline + unknown acceptor -> betaine acetaldehyde + reduced acceptor.
- Glucose dehydrogenase (GLD) (EC 1.1.99.10) from Drosophila. Reaction catalyzed: glucose + unknown acceptor -> delta-gluconolactone + reduced acceptor.
- Cholesterol oxidase (CHOD) (EC 1.1.3.6) from Brevibacterium sterolicum and Streptomyces strain SA-COO. Reaction catalyzed: cholesterol + oxygen -> cholest-4-en-3-one + hydrogen peroxide.
- AlkJ [3], an alcohol dehydrogenase from Pseudomonas oleovorans, which converts aliphatic medium-chain-length alcohols into aldehydes. This family also includes a lyase:
- (R)-mandelonitrile lyase (EC 4.1.2.10) (hydroxynitrile lyase) from plants [4], an enzyme involved in cyanogenis, the release of hydrogen cyanide from injured tissues.

These enzymes are proteins of size ranging from 556 (CHD) to 664 (MOX) amino acid residues which share a number of regions of sequence similarities. One of these regions, located in the N-terminal section, corresponds to the FAD ADP- binding domain. The function of the other conserved domains is not yet known; two of these domains have been selected as signature patterns. The first one is located in the N-terminal section of these enzymes, about 50 residues after the ADP-binding domain, while the second one is located in the central section.

-Consensus pattern: [GA]-[RKN]-x-[LIV]-G(2)-[GST](2)-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-N-x(3)-~~[FYWA]~~[FYWA (SEQ ID NO: 141)]- x(2)-[PAG]-x(5)-~~[DNESH]~~[DNESH (SEQ ID NO: 96)]

-Consensus pattern: [GS]-~~[PSTA]~~[PSTA (SEQ ID NO: 616)]-x(2)-[ST]-P-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-x(2)-S-G-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-G

[ 1] Cavener D.R. J. Mol. Biol. 223:811-814(1992).

[ 2] Henikoff S., Henikoff J.G. Genomics 19:97-107(1994).

[ 3] van Beilen J.B., Eggink G., Enequist H., Bos R., Witholt B. Mol. Microbiol. 6:3121-3136(1992).

[ 4] Cheng I.P., Poulton J.E. Plant Cell Physiol. 34:1139-1143(1993).

77. CKS (Cyclin-dependent kinase regulatory subunit) Number of members: 11. Cyclin-dependent kinases (CDK) are protein kinases which associate with cyclins to regulate eukaryotic cell cycle progression. The most well known CDK is p34-cdc2 (CDC28 in yeast) which is required for entry into S-phase and mitosis. CDK's bind to a regulatory subunit which is essential for their biological function. This regulatory subunit is a small protein of 79 to 150 residues. In yeast (gene CKS1) and in fission yeast (gene suc1) a single isoform is known, while mammals have two highly related isoforms. It has been shown [1] that these CDK regulatory subunits assemble as an hexamer which then acts as a hub for the oligomerization of six CDK catalytic subunits. The sequence of CDK regulatory subunits are highly conserved therefore, the two most conserved regions have been used as signature patterns.

-Consensus pattern: Y-S-x-[KR]-Y-x-[DE](2)-x-[FY]-E-Y-R-H-V-x-[LV]-[PT]-[KRP]
-Consensus pattern: H-x-P-E-x-H-[IV]-L-L-F-[KR]

[ 1] Parge H.E., Arvai A.S., Murtari D.J., Reed S.I., Tainer J.A. Science 262:387-395(1993).

78. CK_II_beta (Casein kinase II regulatory subunit)
Number of members: 16. Casein kinase II (CK-2) [1] is an ubiquitous eukaryotic serine/threonine protein kinase which is found both in the cytoplasm and the nucleus and whose substrates are numerous. It generally phosphorylates Ser or Thr at the N-terminal of stretch of acidic residues (see <PDOC00006>). CK-2 exists as an heterotetramer composed of two catalytic subunits (alpha) and two regulatory subunits (beta). In most species there are two closely related isoforms of the catalytic subunit: alpha and alpha'. Some species, such as fungi and plants, express two forms of regulatory subunits: beta and beta'. The exact function of the regulatory subunit is not yet known. It is a highly conserved protein of about 25 Kd that contains, in its central section, a cysteine-rich motif that could

be involved in binding a metal such as zinc [2]. This region has been used as a signature pattern.

-Consensus pattern: C-P-x-~~[LIVMY]~~[LIVMY (SEQ ID NO: 526)]-x-C-x(5)-[LI]-P-~~[LIVMC]~~[LIVMC (SEQ ID NO: 396)]-G-x(9)-V-[KR]-x(2)-C-P-x-C

[ 1] Allende J.E., Allende C.C. FASEB J. 9:313-323(1995).

[ 2] Reed J.C., Bidwai A.P., Glover C.V.C. J. Biol. Chem. 269:18192-18200(1994).

79. CLP_protease (Clp protease)

These proteins belong to family S14 in the classification of peptidases.

-!- The Clp protease has an active site catalytic triad. In E. coli Clp protease, ser-111, his-136 and asp-185 form the catalytic triad.

-!- Swiss:P48254 has lost all of these active site residues and is therefore inactive.

-!- Swiss:P42379 contains two large insertions, Swiss:P42380 contains one large insertion.

Number of members: 38

The endopeptidase Clp (EC 3.4.21.92) from Escherichia coli cleaves peptides in various proteins in a process that requires ATP hydrolysis [1,2]. Clp is a dimeric protein which consists of a proteolytic subunit (gene clpP) and either of two related ATP-binding regulatory subunits (genes clpA and clpX). ClpP is a serine protease which has a chymotrypsin-like activity. Its catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. Proteases highly similar to ClpP have been found to be encoded in the genome of the chloroplast of plants and seem to be also present in other eukaryotes. The sequences around two of the residues involved in the catalytic triad (a serine and a histidine) are highly conserved and can be used as signature patterns specific to that category of proteases.

-Consensus pattern: T-x(2)-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-G-x-A-[SAC]-S-[MSA]-[PAG]-[STA] [S is the active site residue]

-Consensus pattern: R-x(3)-[EAP]-x(3)-~~[LIVMFYT]~~[LIVMFYT (SEQ ID NO: 460)]-M-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-H-Q-P [H is the active site residue]

[1]Medline: 98050920. The structure of ClpP at 2.3 angstroms resolution suggests a model for ATP-dependent proteolysis. Wang J, Hartling JA, Flanagan JM; Cell 1997;91:447-456.

[ 1] Maurizi M.R., Clark W.P., Kim S.-H., Gottesman S. J. Biol. Chem. 265:12546-12552(1990).

[ 2] Gottesman S., Maurizi M.R. Microbiol. Rev. 56:592-621(1992).

[ 3] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

80. CNG_membrane (Transmembrane region cyclic Nucleotide Gated Channel)

[1]Medline: 94224763. Cyclic nucleotide-gated channels: an expanding new family of ion channels. Yau KW; Proc Natl Acad Sci USA 1994;91:3481-3483.

This family is found to the N-terminus of the cNMP_binding. Number of members: 56. Proteins that bind cyclic nucleotides (cAMP or cGMP) share a structural domain of about 120 residues [1-3]. The best studied of these proteins is the prokaryotic catabolite gene activator (also known as the cAMP receptor protein) (gene crp) where such a domain is known to be composed of three alpha-helices and a distinctive eight-stranded, antiparallel beta-barrel structure. Such a domain is known to exist in the following proteins:

- Prokaryotic catabolite gene activator protein (CAP).
- cAMP- and cGMP-dependent protein kinases (cAPK and cGPK). Both types of kinases contains two tandem copies of the cyclic nucleotide-binding domain. The cAPK's are composed of two different subunits: a catalytic chain and a regulatory chain which contains both copies of the domain. The cGPK's are single chain enzymes that include the two copies of the domain in their N-terminal section. The nucleotide specificity of cAPK and cGPK is due to an amino acid in the conserved region of beta-barrel 7: a threonine that is invariant in cGPK is an alanine in most cAPK.
- Vertebrate cyclic nucleotide-gated ion-channels. Two such cations channels have been fully characterized. One is found in rod cells where it plays a role in visual signal transduction. It specifically binds to cGMP leading to an opening of the channel and thereby causing a depolarization of rod photoreceptors. In olfactory epithelium a similar, cAMP-binding, channel plays a role in odorant signal transduction. There are six invariant amino acids in this domain, three of which are glycine residues that are thought to be essential for maintenance of the structural integrity of the beta-barrel. Two signature patterns have been developed for this domain. The first pattern is located within beta-barrels and 3 and contains the first two conserved Gly. The second pattern is located within beta-

barrels 6 and 7 and contains the third conserved Gly as well as the three other invariant residues.

-Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[VIC]-x(2)-G-~~[DENQTA]~~[DENQTA (SEQ ID NO: 55)]-x-[GAC]-x(2)-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)](4)-x(2)-G

-Consensus pattern: ~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-G-E-x-[GAS]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(5,11)-R-~~[STAQ]~~[STAQ (SEQ ID NO: 730)]-A-x-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)]-x-~~[STACV]~~[STACV (SEQ ID NO: 686)]

[ 1] Weber I.T., Shabb J.B., Corbin J.D. Biochemistry 28:6122-6127(1989).

[ 2] Kaupp U.B. Trends Neurosci. 14:150-157(1991).

[ 3] Shabb J.B., Corbin J.D. J. Biol. Chem. 267:5723-5726(1992).

81. COX10_ctaB_cyoE (Cytochrome c oxidase assembly factor)

[1]Medline: 95191390

Biosynthesis and functional role of haem O and haem A

Mogi T, Saiki K, Anraku Y; Mol Microbiol 1994;14:391-398.

Cytochrome c oxidase is a multi subunit enzyme. The complexity of this enzyme requires assistance in building the complex. This is carried out by the Cytochrome c oxidase assembly factor.

Number of members: 31

Cytochrome c oxidase is an oligomeric enzymatic complex which seems to require the aid of a number of proteins that either act as chaperonins to help the subunits of the enzyme to fold correctly, or assist in the assembly of the metal centers [1]. One of these subunits is known as COX10 in yeast and as ctaB [2] in aerobic prokaryotes. It is evolutionary related to cyoE protein from the Escherichia coli cytochrome O terminal oxidase complex.

These proteins probably contain [3] seven transmembrane segments. The most conserved region is located in a loop between the second and third of these segments and has been selected as a signature pattern.

-Consensus pattern: [ED]-x-D-x(2)-M-x-R-T-x(2)-R-x(4)-G

[ 1] Nobrega M.P., Nobrega F.G., Tzagoloff A.
J. Biol. Chem. 265:14220-14226(1990).
[ 2] Cao J., Hosler J., Shapleigh J., Revzin A., Ferguson-Miller S.
J. Biol. Chem. 267:24273-24278(1992).
[ 3] Chepuri V., Gennis R.B.
J. Biol. Chem. 265:12978-12986(1990).

82. COX3 (Cytochrome c oxidase subunit III)
This family corresponds to chains c and p.
[1]Medline: 96216288
The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S; Science 1996;272:1136-1144.
Number of members: 224

83. COX5B (Cytochrome c oxidase subunit Vb)
[1]
Medline: 96216288
The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A.
Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S;
Science 1996;272:1136-1144.
This family consists of chains F and S
Number of members: 10

Cytochrome c oxidase (EC 1.9.3.1) [1] is an oligomeric enzymatic complex which is a component of the respiratory chain complex and is involved in the transfer of electrons from cytochrome c to oxygen. In eukaryotes this enzyme

complex is located in the mitochondrial inner membrane; in aerobic prokaryotes it is found in the plasma membrane. In addition to the three large subunits that form the catalytic center of the enzyme complex there are, in eukaryotes, a variable number of small polypeptidic subunits. One of these subunits which is known as Vb in mammals, V in slime mold and IV in yeast, binds a zinc atom. The sequence of subunit Vb is well conserved and includes three conserved cysteines that are thought to coordinate the zinc ion [2]. Two of these cysteines are clustered in the C-terminal section of the subunit; this region has been selected as a signature pattern.

-Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-[FYW]-x(10)-C-x(2)-C-G-x(2)-[FY]-K-L [The two C's probably bind zinc]

[ 1] Capaldi R.A., Malatesta F., Darley-Usmar V.M.
Biochim. Biophys. Acta 726:135-148(1983).
[ 2] Rizzuto R., Sandona D., Brini M., Capaldi R.A., Bisson R.
Biochim. Biophys. Acta 1129:100-104(1991).

84. COesterase (Carboxylesterases)
Cholinesterase pages
The prints entry is specific to acetylcholinesterase
Number of members: 273

Higher eukaryotes have many distinct esterases. Among the different types are those which act on carboxylic esters (EC 3.1.1.-). Carboxyl-esterases have been classified into three categories (A, B and C) on the basis of differential patterns of inhibition by organophosphates. The sequence of a number of type-B carboxylesterases indicates [1,2,3] that the majority are evolutionary related. This family currently consists of the following proteins:

- Acetylcholinesterase (EC 3.1.1.7) (AChE) [E1] from vertebrates and from Drosophila.

- Mammalian cholinesterase II (butyryl cholinesterase) (EC 3.1.1.8). Acetylcholinesterase and cholinesterase II are closely related enzymes that hydrolyze choline esters [4].
- Mammalian liver microsomal carboxylesterases (EC 3.1.1.1).
- Drosophila esterase 6, produced in the anterior ejaculatory duct of the male insect reproductive system where it plays an important role in its reproductive biology.
- Drosophila esterase P.
- Culex pipiens (mosquito) esterases B1 and B2.
- Myzus persicae (peach-potato aphid) esterases E4 and FE4.
- Mammalian bile-salt-activated lipase (BAL) [5], a multifunctional lipase which catalyzes fat and vitamin absorption. It is activated by bile salts in infant intestine where it helps to digest milk fats.
- Insect juvenile hormone esterase (JH esterase) (EC 3.1.1.59).
- Lipases (EC 3.1.1.3) from the fungi Geotrichum candidum and Candida rugosa.
- Caenorhabditis gut esterase (gene ges-1).
- Duck fatty acyl-CoA hydrolase, medium chain (EC 3.1.2.14), an enzyme that may be associated with peroxisome proliferation and may play a role in the production of 3-hydroxy fatty acid diester pheromones.
- Membrane enclosed crystal proteins from slime mold. These proteins are, most probably esterases; the vesicles where they are found have therefore been termed esterosomes.

So far two bacterial proteins have been found to belong to this family:

- Phenmedipham hydrolase (phenylcarbamate hydrolase), an Arthrobacter oxidans plasmid-encoded enzyme (gene pcd) that degrades the phenylcarbamate herbicides phenmedipham and desmedipham by hydrolyzing their central carbamate linkages.
- Para-nitrobenzyl esterase from Bacillus subtilis (gene pnbA).

The following proteins, while having lost their catalytic activity, contain a domain evolutionary related to that of carboxylesterases type-B:

- Thyroglobulin (TG), a glycoprotein specific to the thyroid gland, which is the precursor of the iodinated thyroid hormones thyroxine (T4) and triiodo thyronine (T3).
- Drosophila protein neuractin (gene nrt) which may mediate or modulate cell adhesion between embryonic cells during development.
- Drosophila protein glutactin (gene glt), whose function is not known.

As is the case for lipases and serine proteases, the catalytic apparatus of esterases involves three residues (catalytic triad): a serine, a glutamate or aspartate and a histidine. The sequence around the active site serine is well conserved and can be used as a signature pattern. A conserved region located in the N-terminal section containing a cysteine involved in a disulfide bond has been selected as a second signature pattern.

-Consensus pattern: F-[GR]-G-x(4)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-[LIV]-x-G-x-S-~~[STAG]~~[STAG (SEQ ID NO: 690)]-G[S is the active site residue]
-Consensus pattern: [ED]-D-C-L-[YT]-[LIV]-[DNS]-[LIV]-~~[LIVFYW]~~[LIVFYW (SEQ ID NO: 376)]-x-[PQR] [C is involved in a disulfide bond]

[ 1] Myers M., Richmond R.C., Oakeshott J.G. Mol. Biol. Evol. 5:113-119(1988).
[ 2] Krejci E., Duval N., Chatonnet A., Vincens P., Massoulie J. Proc. Natl. Acad. Sci. U.S.A. 88:6647-6651(1991).
[ 3] Cygler M., Schrag J.D., Sussman J.L., Harel M., Silman I. Gentry M.K., Doctor B.P. Protein Sci. 2:366-382(1993).
[ 4] Lockridge O. BioEssays 9:125-128(1988).
[ 5] Wang C.-S., Hartsuck J.A. Biochim. Biophys. Acta 1166:1-19(1993).

85. CPSase_L_chain (Carbamoyl-phosphate synthase (CPSase))

[1]

Medline: 94347758
Three-dimensional structure of the biotin carboxylase subunit.
of acetyl-CoA carboxylase.
Waldrop GL, Rayment I, Holden HM;

Biochemistry 1994;33:10249-10256.

[1]

Medline: 90285162

Mammalian carbamyl phosphate synthetase (CPS). DNA sequence and evolution of the CPS domain of the Syrian hamster multifunctional protein CAD.

Simmer JP, Kelly RE, Rinker AG Jr, Scully JL, Evans DR;

Biol Chem 1990;265:10395-10402.

Carbamoyl-phosphate synthase catalyzes the ATP-dependent synthesis of carbamyl-phosphate from glutamine or ammonia and bicarbonate. This important enzyme initiates both the urea cycle and the biosynthesis of arginine and/or pyrimidines [2].

The carbamoyl-phosphate synthase (CPS) enzyme in prokaryotes is a heterodimer of a small and large chain. The small chain promotes the hydrolysis of glutamine to ammonia, which is used by the large chain to synthesize carbamoyl phosphate. See CPSase_sm_chain.

The small chain has a GATase domain in the carboxyl terminus. See GATase.

Number of members: 181

Carbamoyl-phosphate synthase (CPSase) catalyzes the ATP-dependent synthesis of carbamyl-phosphate from glutamine (EC 6.3.5.5) or ammonia (EC 6.3.4.16) and bicarbonate [1]. This important enzyme initiates both the urea cycle and the biosynthesis of arginine and pyrimidines.

Glutamine-dependent CPSase (CPSase II) is involved in the biosynthesis of pyrimidines and purines. In bacteria such as Escherichia coli, a single enzyme is involved in both biosynthetic pathways while other bacteria have separate enzymes. The bacterial enzymes are formed of two subunits. A small chain (gene carA) that provides glutamine amidotransferase activity (GATase) necessary for removal of the ammonia group from glutamine, and a large chain (gene carB) that provides CPSase activity. Such a structure is also present in fungi for arginine biosynthesis (genes CPA1 and CPA2). In most eukaryotes, the first three steps of pyrimidine biosynthesis are catalyzed by a large

multifunctional enzyme - called URA2 in yeast, rudimentary in Drosophila and CAD in mammals [2]. The CPSase domain is located between an N-terminal GATase domain and the C-terminal part which encompass the dihydroorotase and aspartate transcarbamylase activities.

Ammonia-dependent CPSase (CPSase I) is involved in the urea cycle in ureolytic vertebrates; it is a monofunctional protein located in the mitochondrial matrix.

The CPSase domain is typically 120 Kd in size and has arisen from the duplication of an ancestral subdomain of about 500 amino acids. Each subdomain independently binds to ATP and it is suggested that the two homologous halves act separately, one to catalyze the phosphorylation of bicarbonate to carboxy phosphate and the other that of carbamate to carbamyl phosphate.

The CPSase subdomain is also present in a single copy in the biotin-dependent enzymes acetyl-CoA carboxylase (EC 6.4.1.2) (ACC), propionyl-CoA carboxylase (EC 6.4.1.3) (PCCase), pyruvate carboxylase (EC 6.4.1.1) (PC) and urea carboxylase (EC 6.3.4.6).

Two conserved regions which are probably important for binding ATP and/or catalytic activity have been selected as signatures for the subdomain.

-Consensus pattern: [FYV]-[PS]-~~[LIVMC]~~[LIVMC (SEQ ID NO: 396)]-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[KR]-[PSA]-[STA]-x(3)-[SG]-G-x-[AG]

-Consensus pattern: ~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-~~[LIMN]~~[LIMN (SEQ ID NO: 342)]-E-~~[LIVMCA]~~[LIVMCA (SEQ ID NO: 397)]-N-~~[PATLIVM]~~[PATLIVM (SEQ ID NO: 595)]-[KR]-~~[LIVMSTAC]~~[LIVMSTAC (SEQ ID NO: 511)]

[ 1] Simmer J.P., Kelly R.E., Rinker A.G. Jr., Scully J.L., Evans D.R. J. Biol. Chem. 265:10395-10402(1990).

[ 2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-164(1993).

86. CPSase_sm_chain (Carbamoyl-phosphate synthase small chain, CPSase domain) [1]

Medline: 90285162

Mammalian carbamyl phosphate synthetase (CPS). DNA sequence and evolution of the CPS domain of the Syrian hamster multifunctional protein CAD.

Simmer JP, Kelly RE, Rinker AG Jr, Scully JL, Evans DR;

Biol Chem 1990;265:10395-10402.

The carbamoyl-phosphate synthase domain is in the amino terminus of protein.

Carbamoyl-phosphate synthase catalyzes the ATP-dependent synthesis of carbamyl-phosphate from glutamine or ammonia and bicarbonate. This important enzyme initiates both the urea cycle and the biosynthesis of arginine and/or pyrimidines [1].

The carbamoyl-phosphate synthase (CPS) enzyme in prokaryotes is a heterodimer of a small and large chain. The small chain promotes the hydrolysis of glutamine to ammonia, which is used by the large chain to synthesize carbamoyl phosphate. See CPSase_L_chain.

The small chain has a GATase domain in the carboxyl terminus. See GATase.

Number of members: 46

Carbamoyl-phosphate synthase (CPSase) catalyzes the ATP-dependent synthesis of carbamyl-phosphate from glutamine (EC 6.3.5.5) or ammonia (EC 6.3.4.16) and bicarbonate [1]. This important enzyme initiates both the urea cycle and the biosynthesis of arginine and pyrimidines.

Glutamine-dependent CPSase (CPSase II) is involved in the biosynthesis of pyrimidines and purines. In bacteria such as Escherichia coli, a single enzyme is involved in both biosynthetic pathways while other bacteria have separate enzymes. The bacterial enzymes are formed of two subunits. A small chain (gene carA) that provides glutamine amidotransferase activity (GATase) necessary for

removal of the ammonia group from glutamine, and a large chain (gene carB) that provides CPSase activity. Such a structure is also present in fungi for arginine biosynthesis (genes CPA1 and CPA2). In most eukaryotes, the first three steps of pyrimidine biosynthesis are catalyzed by a large multifunctional enzyme - called URA2 in yeast, rudimentary in Drosophila and CAD in mammals [2]. The CPSase domain is located between an N-terminal GATase domain and the C-terminal part which encompass the dihydroorotase and aspartate transcarbamylase activities.

Ammonia-dependent CPSase (CPSase I) is involved in the urea cycle in ureolytic vertebrates; it is a monofunctional protein located in the mitochondrial matrix.

The CPSase domain is typically 120 Kd in size and has arisen from the duplication of an ancestral subdomain of about 500 amino acids. Each subdomain independently binds to ATP and it is suggested that the two homologous halves act separately, one to catalyze the phosphorylation of bicarbonate to carboxy phosphate and the other that of carbamate to carbamyl phosphate.

The CPSase subdomain is also present in a single copy in the biotin-dependent enzymes acetyl-CoA carboxylase (EC 6.4.1.2) (ACC), propionyl-CoA carboxylase (EC 6.4.1.3) (PCCase), pyruvate carboxylase (EC 6.4.1.1) (PC) and urea carboxylase (EC 6.3.4.6).

Two conserved regions which are probably important for binding ATP and/or catalytic activity have been selected as signatures for the subdomain.

-Consensus pattern: [FYV]-[PS]-~~[LIVMC]~~[LIVMC (SEQ ID NO: 396)]-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[KR]-[PSA]-[STA]-x(3)-[SG]-G-x-[AG]

-Consensus pattern: ~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-~~[LIMN]~~[LIMN (SEQ ID NO: 342)]-E-~~[LIVMCA]~~[LIVMCA (SEQ ID NO: 397)]-N-~~[PATLIVM]~~[PATLIVM (SEQ ID NO: 595)]-[KR]-~~[LIVMSTAC]~~[LIVMSTAC (SEQ ID NO: 511)]

[ 1] Simmer J.P., Kelly R.E., Rinker A.G. Jr., Scully J.L., Evans D.R.
J. Biol. Chem. 265:10395-10402(1990).

[ 2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B.
BioEssays 15:157-164(1993).

87. CRAL_TRIO (CRAL/TRIO domain)
[1]
Medline: 98121119
Crystal structure of the Saccharomyces cerevisiae phosphatidyl-inositol-transfer protein.
Sha B, Phillips SE, Bankaitis VA, Luo M;
Nature 1998;391:506-510.
The original profile has been extended to include the carboxyl domain from the known structure of Sec14. Swiss:P10911 has not been included in the Pfam family because it does not appear to contain a complete structural domain.
Number of members: 39

88. CSD ( 'Cold-shock' DNA-binding domain)
[1]
Medline: 94255482
Crystal structure of CspA, the major cold shock protein of Escherichia coli.
Schindelin H, Jiang W, Inouye M, Heinemann U;
Proc Natl Acad Sci U S A 1994;91:5119-5123.
Number of members: 121

A conserved domain of about 70 amino acids has been found in prokaryotic and eukaryotic DNA-binding proteins [1,2,3,E1]. This domain, which is known as the 'cold-shock domain' (CSD) is present in the proteins listed below.

- Escherichia coli protein CS7.4 (gene cspA) which is induced in response to

low temperature (cold-shock protein) and which binds to and stimulates the transcription of the CCAAT-containing promoters of the HN-S protein and of gyrA.

- Mammalian Y box binding protein 1 (YB1). A protein that binds to the CCAAT-containing Y box of mammalian HLA class II genes.
- Xenopus Y box binding proteins -1 and -2 (Y1 and Y2). Proteins that bind to the CCAAT-containing Y box of Xenopus hsp70 genes.
- Xenopus B box binding protein (YB3). YB3 binds the B box promoter element of genes transcribed by RNA polymerase III.
- Enhancer factor I subunit A (EFI-A) (dbpB). A protein that also bind to CCAAT-motif in various gene promoters.
- DbpA, a Human DNA-binding protein of unknown specificity.
- Bacillus subtilis cold-shock proteins cspB and cspC.
- Streptomyces clavuligerus protein SC 7.0.
- Escherichia coli proteins cspB, cspC, cspD, cspE and cspF.
- Unr, a mammalian gene encoded upstream of the N-ras gene. Unr contains nine repeats that are similar to the CSD domain. The function of Unr is not yet known but it could be a multivalent DNA-binding protein.

As a signature pattern for the CSD domain, its most conserved region which is located in its N-terminal section has been selected. It must be noted that the beginning of this region is highly similar [4] to the RNP-1 RNA-binding motif.

-Consensus pattern: [FY]-G-F-I-x(6,7)-[DER]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-F-x-H-x-~~[STKR]~~[STKR (SEQ ID NO: 752)]-x-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]

[ 1] Doniger J., Landsman D., Gonda M.A., Wistow G.
New Biol. 4:389-395(1992).

[ 2] Wistow G.
Nature 344:823-824(1990).

[ 3] Jones P.G., Inouye M.
Mol. Microbiol. 11:811-818(1994).

[ 4] Landsman D.
Nucleic Acids Res. 20:2861-2864(1992).

89. CTF_NFI (CTF/NF-I family)
Number of members: 45

Nuclear factor I (NF-I) or CCAAT box-binding transcription factor (CTF) [1,2] (also known as TGGCA-binding proteins) are a family of vertebrate nuclear proteins which recognize and bind, as dimers, the palindromic DNA sequence 5'-TGGCANNNTGCCA-3'. CTF/NF-I binding sites are present in viral and cellular promoters and in the origin of DNA replication of Adenovirus type 2.

The CTF/NF-I proteins were first identified as nuclear factor I, a collection of proteins that activate the replication of several Adenovirus serotypes (together with NF-II and NF-III) [3]. The family of proteins was also identified as the CTF transcription factors, before the NFI and CTF families were found to be identical [4]. The CTF/NF-I proteins are individually capable of activating transcription and DNA replication. The CTF/NF-I family name has also been dubbed as NFI, NF-I or NF1.

In a given species, there are a large number of different CTF/NF-I proteins. The multiplicity of CTF/NF-I is known to be generated both by alternative splicing and by the occurrence of four different genes. The known forms of NF-I genes have been classified as:

- The CTF-like factors subfamily (prototype form: CTF-1) [4]
- The NFI-X proteins.
- The NFI-A proteins.
- The NFI-B proteins.

So far, all CTF/NF-I family members appear to have similar transcription and replication activities.

CTF/NF-1 proteins contains 400 to 600 amino acids. The N-terminal 200 amino-acid sequence, almost perfectly conserved in all species and genes sequenced,

mediates site-specific DNA recognition, protein dimerization and Adenovirus DNA replication. The C-terminal 100 amino acids contain the transcriptional activation domain. This activation domain is the target of gene expression regulatory pathways ellicited by growth factors and it interacts with basal transcription factors and with histone H3 [6].

A perfectly conserved, highly charged 12 residue peptide located in the N-terminal part of CTF/NF-I has been selected as a specific signature for this family of proteins.

-Consensus pattern: R-K-R-K-Y-F-K-K-H-E-K-R

[ 1] Mermod N., O'Neill E.A., Kelly T.J., Tjian R.
Cell 58:741-753(1989).
[ 2] Rupp R.A.W., Kruse U., Multhaup G., Goebel U., Beyreuther K., Sippel A.E.
Nucleic Acids Res. 18:2607-2616(1990).
[ 3] Nagata K., Guggenheimer R.A., Enomoto T., Lichy J.H., Hurwitz J.
Proc. Natl. Acad. Sci. U.S.A. 79:6438-6442(1982).
[ 4] Santoro C., Mermod N., Andrews P.C., Tjian R.
Nature 334:2118-2224(1988).
[ 5] Gil G., Smith J.R., Goldstein J.L., Slaughter C.A., Orth K., Brown M.S., Osborne T.F.
Proc. Natl. Acad. Sci. U.S.A 85:8963-8967(1988).
[ 6] Alevizopoulos A., Dusserre Y., Tsai-Pflugfelder M., von der Weid T., Wahli W., Mermod N.
Genes Dev. 9:3051-3066(1995).

90. Calsequestrin (Calsequestrin)
Number of members: 13

Calsequestrin is a moderate-affinity, high-capacity calcium-binding protein of cardiac and skeletal muscle [1], where it is located in the lumenal space of the sarcoplasmic reticulum terminal cisternae. Calsequestrin acts as a

calcium buffer and plays an important role in the muscle excitation-contraction coupling. It is a highly acidic protein of about 400 amino acid residues that binds more than 40 moles of calcium per mole of protein. There are at least two different forms of calsequestrin: one which is expressed in cardiac muscles and another in skeletal muscles. Both forms have highly similar sequences.

Two signature sequences have been developed. The first corresponds to the N-terminus of the mature protein, the second is located just in front of the C-terminus of the protein which is composed of a highly acidic tail of variable length.

-Consensus pattern: [EQ]-[DE]-G-L-[DN]-F-P-x-Y-D-G-x-D-R-V

-Consensus pattern: [DE]-L-E-D-W-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-E-D-V-L-x-G-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-N-T-E-D-D-D

[ 1] Treves S., Vilsen B., Chiozzi P., Andersen J.P., Zorzato F.
Biochem. J. 283:767-772(1992).

91. Carboxyl_trans (Carboxyl transferase domain)

[1]

Medline: 93374821
Primary structure of the monomer of the 12S subunit of transcarboxylase as deduced from DNA and characterization of the product expressed in Escherichia coli.
Thornton CG, Kumar GK, Haase FC, Phillips NF, Woo SB, Park VM, Magner WJ, Shenoy BC, Wood HG, Samols D;
J Bacteriol 1993;175:5301-5308.

[2]

Medline: 93358891
Molecular evolution of biotin-dependent carboxylases.
Toh H, Kondo H, Tanabe T;
Eur J Biochem 1993;215:687-696.

All of the members in this family are biotin dependent carboxylases. The carboxyl transferase domain carries out the following reaction; transcarboxylation from biotin to an acceptor molecule. There are two recognised types of carboxyl transferase. One of them uses acyl-CoA and the other uses 2-oxo acid as the acceptor molecule of carbon dioxide. All of the members in this family utilise acyl-CoA as the acceptor molecule.

Number of members: 47

92. Chal_stil_synt (Chalcone and stilbene synthases)

Number of members: 146

Chalcone synthases (CHS) (EC 2.3.1.74) and stilbene synthases (STS) (formerly known as resveratrol synthases) are related plant enzymes [1]. CHS is an important enzyme in flavanoid biosynthesis and STS a key enzyme in stilbene-type phyloalexin biosynthesis. Both enzymes catalyze the addition of three molecules of malonyl-CoA to a starter CoA ester (a typical example is 4-coumaroyl-CoA), producing either a chalcone (with CHS) or stilbene (with STS).

These enzymes are proteins of about 390 amino-acid residues. A conserved cysteine residue, located in the central section of these proteins, has been shown [2] to be essential for the catalytic activity of both enzymes and probably represents the binding site for the 4-coumaryl-CoA group. The region around this active site residue is well conserved and can be used as a signature pattern.

In addition to the plant enzymes, this family also includes Bacillus subtilis bcsA.

-Consensus pattern: R-~~[LIVMFYS]~~[LIVMFYS (SEQ ID NO: 452)]-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-[QHG]-x-G-C-~~[FYNA]~~[FYNA (SEQ ID NO: 134)]-[GA]-G-[GA]-

~~[STAV]~~[STAV (SEQ ID NO: 733)]-x-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-[RA] [C is the active site residue]

[ 1] Schroeder J., Schroeder G.
Z. Naturforsch. 45C:1-8(1990).
[ 2] Lanz T., Tropf S., Marner F.-J., Schroeder J., Schroeder G.
J. Biol. Chem. 266:9971-9976(1991).

93. Chorismate_synt (Chorismate synthase)
Number of members: 19

Chorismate synthase (EC 4.6.1.4) catalyzes the last of the seven steps in the shikimate pathway which is used in prokaryotes, fungi and plants for the biosynthesis of aromatic amino acids. It catalyzes the 1,4-trans elimination of the phosphate group from 5-enolpyruvylshikimate-3-phosphate (EPSP) to form chorismate which can then be used in phenylalanine, tyrosine or tryptophan biosynthesis. Chorismate synthase requires the presence of a reduced flavin mononucleotide (FMNH2 or FADH2) for its activity.

Chorismate synthase from various sources shows [1,2] a high degree of sequence conservation. It is a protein of about 360 to 400 amino-acid residues.
Three signature patterns have been developed from conserved regions rich in basic residues (mostly arginines). The first is in the N-terminal section, the second is central and the third is C-terminal.

-Consensus pattern: G-E-S-H-[GC]-x(2)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[GTV]-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-[DE]-G-x-[PV]

-Consensus pattern: [GE]-R-[SA](2)-[SAG]-R-[EV]-[ST]-x(2)-[RH]-V-x(2)-G

-Consensus pattern: R-[SH]-D-[PSV]-~~[CSAV]~~[CSAV (SEQ ID NO: 16)]-x(4)-[GAI]-x-~~[IVGSP]~~[IVGSP (SEQ ID NO: 279)]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-E-~~[STAH]~~[STAH (SEQ ID NO: 711)]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]

[1] Schaller A., Schmid J., Leibinger U., Amrhein N.
J. Biol. Chem. 266:21434-21438(1991).
[2] Jones D.G.L., Reusser U., Braus G.H.
Mol. Microbiol. 5:2143-2152(1991).

94. Clat_adaptor_s (Clathrin adaptor complex small chain)
Number of members: 21

Clathrin coated vesicles (CCV) mediate intracellular membrane traffic such as receptor mediated endocytosis. In addition to clathrin, the CCV are composed of a number of other components including oligomeric complexes which are known as adaptor or clathrin assembly proteins (AP) complexes [1]. The adaptor complexes are believed to interact with the cytoplasmic tails of membrane proteins, leading to their selection and concentration. In mammals two type of adaptor complexes are known: AP-1 which is associated with the Golgi complex and AP-2 which is associated with the plasma membrane. Both AP-1 and AP-2 are heterotetramers that consist of two large chains - the adaptins - (gamma and beta' in AP-1; alpha and beta in AP-2); a medium chain (AP47 in AP-1; AP50 in AP-2) and a small chain (AP19 in AP-1; AP17 in AP-2).

The small chains of AP-1 and AP-2 are evolutionary related proteins of about 18 Kd. Homologs of AP17 and AP19 have also been found in yeast (genes APS1/YAP19 and APS2/YAP17) [2,3,4]. AP17 and AP19 are also related to the zeta-chain [5] of coatomer (zeta-cop), a cytosolic protein complex that reversibly associates with Golgi membranes to form vesicles that mediate biosynthetic protein transport from the endoplasmic reticulum, via the Golgi up to the trans Golgi network.

A conserved region in the central section of these proteins has been selected as a signature pattern.

-Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-Y-[KR]-x(4)-L-Y-F

[ 1] Pearse B.M., Robinson M.S.
Annu. Rev. Cell Biol. 6:151-171(1990).

[ 2] Kirchhausen T., Davis A.C., Frucht S., O'Brine Greco B., Payne G.S., Tubb B.
J. Biol. Chem. 266:11153-11157(1991).

[ 3] Nakai M., Takada T., Endo T.
Biochim. Biophys. Acta 1174:282-284(1993).

[ 4] Phan H.L., Finlay J.A., Chu D.S., Tan P.K., Kirchhausen T., Payne G.S.
EMBO J. 13:1706-1717(1994).

[ 5] Kuge O., Hara-Kuge S., Orci L., Ravazzola M., Amherdt M., Tanigawa G., Wieland F.T., Rothman J.E.
J. Cell Biol. 123:1727-1734(1993).

95. Clathrin_lg_ch (Clathrin light chain.)
Number of members: 8

Clathrin [1,2] is the major coat-forming protein that encloses vesicles such as coated pits and forms cell surface patches involved in membrane traffic within eukaryotic cells. The clathrin coats (called triskelions) are composed of three heavy chains (180 Kd) and three light chains (23 to 27 Kd).

The clathrin light chains [3], which may help to properly orient the assembly and disassembly of the clathrin coats, bind non-covalently to the heavy chain, they also bind calcium and interact with the hsc70 uncoating ATPase.

- In higher eukaryotes two genes code for distinct but related light chains: LC(a) and LC(b). Each of the two genes can yield, by tissue-specific alternative splicing, two separate forms which differ by the insertion of a sequence of respectively thirty or eighteen residues. There is, in the N-terminal part of the clathrin light chains a domain of twenty one amino acid residues which is perfectly conserved in LC(a) and LC(b).
- In yeast there is a single light chain (gene CLC1) whose sequence is only distantly related to that of higher eukaryotes.

Two signature patterns have been developed for clathrin light chains. The first pattern is a heptapeptide from the center of the conserved N-terminal region of eukaryotic light chains; the second pattern is derived from a positively charged region located in the C-terminal extremity of all known clathrin light chains.

-Consensus pattern: F-L-A-Q-Q-E-S

[ 1] Keen J.H.
Annu. Rev. Biochem. 59:415-438(1990).
[ 2] Brodsky F.M.
Science 242:1396-1402(1988).
[ 3] Brodsky F.M., Hill B.L., Acton S.L., Naethke I., Wong D.H., Ponnambalam S., Parham P.
Trends Biochem. Sci. 16:208-213(1991).

96. (Clathrin repeat) 7-fold repeat in Clathrin and VPS
Each repeat is about 140 amino acids long. The repeats occur in the arm region of the Clathrin heavy chain.
Number of members: 79

[1]
Medline: 92191269
Folding and trimerization of clathrin subunits at the triskelion hub.
Nathke IS, Heuser J, Lupas A, Stock J, Turck CW, Brodsky FM;
Cell 1992;68:899-910. [2]
Medline: 88097376
Clathrin heavy chain: molecular cloning and complete primary structure.
Kirchhausen T, Harrison SC, Chow EP, Mattaliano RJ, Ramachandran KL, Smart J, Brosius J;
Proc Natl Acad Sci U S A 1987;84:8805-8809.

97. Collagen (Collagen triple helix repeat (20 copies))

[1] Medline: 94059583

New members of the collagen superfamily

Mayne R, Brewton RG;

Curr Opin Cell Biol 1993;5:883-890.

Scurvy is associated with collagens.

Members of this family belong to the collagen superfamily [1].

Collagens are generally extracellular structural proteins involved in formation of connective tissue structure.

The alignment contains 20 copies of the G-X-Y repeat that forms a triple helix. The first position of the repeat is glycine, the second and third positions can be any residue but are frequently proline and hydroxyproline. Collagens are post translationally modified by proline hydoxylase to form the hydroxyproline residues. Defective hydroxylation is the cause of scurvy.

Some members of the collagen superfamily are not involved in connective tissue structure but share the same triple helical structure.

Number of members: 2125

98. Coprogen_oxidas (Coproporphyrinogen III oxidase)

Number of members: 12

Coproporphyrinogen III oxidase (EC 1.3.3.3) (coproporphyrinogenase) [1,2] catalyzes the oxidative decarboxylation of coproporphyrinogen III into protoporphyrinogen IX, a common step in the pathway for the biosynthesis of porphyrins such as heme, chlorophyll or cobalamin.

Coproporphyrinogen III oxidase is an enzyme that requires iron for its activity. A cysteine seems to be important for the catalytic mechanism [3]. Sequences from a variety of eukaryotic and prokaryotic sources show that

this enzyme has been evolutionarily conserved. A highly conserved region in the central part of the sequence has been selected as a signature pattern. This region contains the only conserved cysteine and is rich in charged amino acids.

-Consensus pattern: K-x-W-C-x(2)-[FYH](3)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-H-R-x-E-x-R-G-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-G-G-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-F-F-D

[ 1] Xu K., Elliott T.
J. Bacteriol. 175:4990-4999(1993).
[ 2] Kohno H., Furukawa T., Yoshinaga T., Tokunaga R., Taketani S.
J. Biol. Chem. 268:21359-21363(1993).
[ 3] Camadro J.M., Chambon H., Jolles J., Labbe P.
Eur. J. Biochem. 156:579-587(1986).
[ 4] Xu K., Elliott T.
J. Bacteriol. 176:3196-3203(1994).

99. Corona_nucleoca (Coronavirus nucleocapsid protein)
[1]
Medline: 98087828
Identification of a specific interaction between the coronavirus mouse hepatitis virus A59 nucleocapsid protein and packaging signal.
Molenkamp R, Spaan WJ;
Virology 1997;239:78-86.
Number of members: 44

100. Cu-oxidase (Multicopper oxidase)
[1]
Medline: 90126844
The blue oxidases, ascorbate oxidase, laccase and ceruloplasmin.

Modelling and structural relationships.

Messerschmidt A, Huber R;

Eur J Biochem 1990;187:341-352.

Number of members: 150

Multicopper oxidases [1,2] are enzymes that possess three spectroscopically different copper centers. These centers are called: type 1 (or blue), type 2 (or normal) and type 3 (or coupled binuclear). The enzymes that belong to this family are:

- Laccase (EC 1.10.3.2) (urishiol oxidase), an enzyme found in fungi and plants, which oxidizes many different types of phenols and diamines.
- Ascorbate oxidase (EC 1.10.3.3), a higher plant enzyme.
- Ceruloplasmin (EC 1.16.3.1) (ferroxidase), a protein found in the serum of mammals and birds, which oxidizes a great variety of inorganic and organic substances. Structurally ceruloplasmin exhibits internal sequence homology, and seem to have evolved from the triplication of a copper-binding domain similar to that found in laccase and ascorbate oxidase.

In addition to the above enzymes there are a number of proteins which, on the basis of sequence similarities, can be said to belong to this family. These proteins are:

- Copper resistance protein A (copA) from a plasmid in Pseudomonas syringae. This protein seems to be involved in the resistance of the microbial host to copper.
- Blood coagulation factor V (Fa V).
- Blood coagulation factor VIII (Fa VIII) [E1].
- Yeast FET3 [3], which is required for ferrous iron uptake.
- Yeast hypothetical protein YFL041w and SpAC1F7.08, the fission yeast homolog.

Factors V and VIII act as cofactors in blood coagulation and are structurally similar [4]. Their sequence consists of a triplicated A domain, a B domain and

a duplicated C domain; in the following order: A-A-B-A-C-C. The A-type domain is related to the multicopper oxidases.

Two signature patterns have been developed for these proteins. Both patterns are derived from the same region, which in ascorbate oxidase, laccase, in the third domain of ceruloplasmin, and in copA, contains five residues that are known to be involved in the binding of copper centers. The first pattern does not make any assumption on the presence of copper-binding residues and thus can detect domains that have lost the ability to bind copper (such as those in Fa V and Fa VIII), while the second pattern is specific to copper-binding domains.

-Consensus pattern: G-x-[FYW]-x-~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)]-x-[CST]-x(8)-G-[LM]-x(3)-~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)]

-Consensus pattern: H-C-H-x(3)-H-x(3)-[AG]-[LM]

[The first two H's are copper type 3 binding residues]

[The C, the 3rd H, and L or M are copper type 1 ligands]

101. Cullin (Cullin family)

Number of members: 24

The following proteins are collectively termed cullins [1]:

- Caenorhabditis elegans cul-1 (or lin-19), a protein required for developmentally programmed transitions from the G1 phase of the cell cycle to the G0 phase or the apoptotic pathway.
- Caenorhabditis elegans cul-2, cul-3, cul-4 (F45E12.3), cul-5 (ZK856.1) and cul-6 (K08E7.7).
- Mammalian CUL1, CUL2, CUL3, CUL4A and CUL4B.
- Mammalian vasopressin-activated calcium-mobilizing receptor (VACM-1), a kidney-specific protein thought to form a cell surface receptor [2] but which does not have any structural hallmarks of a receptor.
- Drosophila lin19.

- Yeast CDC53 [3], which acts in concert with CDC4 and UBC3 (CDC34) to control the G1-to-S phase transition.
- Yeast hypothetical protein YGR003w.
- Fission yeast hypothetical protein SpAC24H6.03.

The cullins are hydrophilic proteins of 740 to 815 amino acids. The C-terminal extremity is the most conserved part of these proteins. A signature pattern has been developed from that region.

-Consensus pattern: [LIV]-K-x(2)-[LIV]-x(2)-L-I-[DEQ]-~~[KRHNQ]~~[KRHNQ (SEQ ID NO: 301)]-x-Y-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-R-x(6,7)-[FY]-x-Y-x-[SA]>

[1] Kipreos E.T., Lander L.E., Wing J.P., He W.W., Hedgecock E.M.
Cell 85:829-839(1996).
[2] Burnatowska-Hledin M.A., Spielman W.S., Smith W.L., Shi P., Meyer J.M., Dewitt D.L.
Am. J. Physiol. 268:f1198-F1210(1995).
[3] Mathias N., Johnson S.L., Winey M., Adams A.E., Goetsch L., Pringle J.R., Byers B., Goebl M.G.
Mol. Cell. Biol. 16:6634-6643(1996).

102. (Cu_amine_oxid)
Copper amine oxidase signatures

Amine oxidases (AO) [1] are enzymes that catalyze the oxidation of a wide range of biogenic amines including many neurotransmitters, histamine and xenobiotic amines. There are two classes of amine oxidases: flavin-containing (EC 1.4.3.4) and copper-containing (EC 1.4.3.6).

Copper-containing AO is found in bacteria, fungi, plants and animals, it is an homodimeric enzyme that binds one copper ion per subunit as well as a 2,4,5- trihydroxyphenylalanine quinone (or topaquinone) (TPQ) cofactor. This cofactor is derived from a tyrosine residue.

Two signature patterns were derived for copper AO, the first one contains the tyrosine which give rises to the TPQ cofactor while the second one contains one of the three histidines that bind the copper atom [2].

Consensus pattern~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x(4)-[ST]-x(2)-N-Y-[DE]-[YN] [The first Y gives rises to TPQ] Sequences known to belong to this class detected by the patternALL.

Consensus patternT-x-[GS]-x(2)-H-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x(3)-E-[DE]-x-P [H is a copper ligand] Sequences known to belong to this class detected by the pattern ALL, except for lentil AO.

[ 1] Knowles P.F., Dooley D.M. (In) Metal ions in biological systems; Sigel H., Sigel A., Eds., 30:361- 403, Marcel Dekker, New-York, (1993).

[ 2] Parsons M.R., Convery M.A., Wilmot C.M., Yadav K.D.S., Blakeley V., Corner A.S., Phillips S.E.V., McPherson M.J., Knowles P.F. Structure 3:1171-1184(1995).

103. Cys-protease (Cysteine protease)

Number of members: 358

Eukaryotic thiol proteases (EC 3.4.22.-) [1] are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. The proteases which are currently known to belong to this family are listed below (references are only provided for recently determined sequences).

- Vertebrate lysosomal cathepsins B (EC 3.4.22.1), H (EC 3.4.22.16), L (EC 3.4.22.15), and S (EC 3.4.22.27) [2].
- Vertebrate lysosomal dipeptidyl peptidase I (EC 3.4.14.1) (also known as cathepsin C) [2].
- Vertebrate calpains (EC 3.4.22.17). Calpains are intracellular calcium-activated thiol protease that contain both a N-terminal catalytic domain

and a C-terminal calcium-binding domain.

- Mammalian cathepsin K, which seems involved in osteoclastic bone resorption [3].
- Human cathepsin O [4].
- Bleomycin hydrolase. An enzyme that catalyzes the inactivation of the antitumor drug BLM (a glycopeptide).
- Plant enzymes: barley aleurain (EC 3.4.22.16), EP-B1/B4; kidney bean EP-C1, rice bean SH-EP; kiwi fruit actinidin (EC 3.4.22.14); papaya latex papain (EC 3.4.22.2), chymopapain (EC 3.4.22.6), caricain (EC 3.4.22.30), and proteinase IV (EC 3.4.22.25); pea turgor-responsive protein 15A; pineapple stem bromelain (EC 3.4.22.32); rape COT44; rice oryzain alpha, beta, and gamma; tomato low-temperature induced, Arabidopsis thaliana A494, RD19A and RD21A.
- House-dust mites allergens DerP1 and EurM1.
- Cathepsin B-like proteinases from the worms Caenorhabditis elegans (genes gcp-1, cpr-3, cpr-4, cpr-5 and cpr-6), Schistosoma mansoni (antigen SM31) and Japonica (antigen SJ31), Haemonchus contortus (genes AC-1 and AC-2), and Ostertagia ostertagi (CP-1 and CP-3).
- Slime mold cysteine proteinases CP1 and CP2.
- Cruzipain from Trypanosoma cruzi and brucei.
- Throphozoite cysteine proteinase (TCP) from various Plasmodium species.
- Proteases from Leishmania mexicana, Theileria annulata and Theileria parva.
- Baculoviruses cathepsin-like enzyme (v-cath).
- Drosophila small optic lobes protein (gene sol), a neuronal protein that contains a calpain-like domain.
- Yeast thiol protease BLH1/YCP1/LAP3.
- Caenorhabditis elegans hypothetical protein C06G4.2, a calpain-like protein.

Two bacterial peptidases are also part of this family:

- Aminopeptidase C from Lactococcus lactis (gene pepC) [5].
- Thiol protease tpr from Porphyromonas gingivalis.

Three other proteins are structurally related to this family, but may have lost their proteolytic activity.

- Soybean oil body protein P34. This protein has its active site cysteine replaced by a glycine.
- Rat testin, a sertoli cell secretory protein highly similar to cathepsin L but with the active site cysteine is replaced by a serine. Rat testin should not be confused with mouse testin which is a LIM-domain protein (see <PDOC00382>).
- Plasmodium falciparum serine-repeat protein (SERA), the major blood stage antigen. This protein of 111 Kd possesses a C-terminal thiol-protease-like domain [6], but the active site cysteine is replaced by a serine.

The sequences around the three active site residues are well conserved and can be used as signature patterns.

-Consensus pattern: Q-x(3)-[GE]-x-C-[YW]-x(2)-~~[STAGC]~~[STAGC (SEQ ID NO: 691)]-~~[STAGCV]~~[STAGCV (SEQ ID NO: 698)] [C is the active site residue]

-Consensus pattern: ~~[LIVMGSTAN]~~[LIVMGSTAN (SEQ ID NO: 492)]-x-H-~~[GSACE]~~[GSACE (SEQ ID NO: 188)]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-~~[LIVMAT]~~[LIVMAT (SEQ ID NO: 394)](2)-G-x-~~[GSADNH]~~[GSADNH (SEQ ID NO: 196)] [H is the active site residue]

-Consensus pattern: ~~[FYCH]~~[FYCH (SEQ ID NO: 124)]-[WI]-~~[LIVT]~~[LIVT (SEQ ID NO: 538)]-x-~~[KRQAG]~~[KRQAG (SEQ ID NO: 318)]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-~~[LFYW]~~[LFYW (SEQ ID NO: 333)]-~~[LIVMFYG]~~[LIVMFYG (SEQ ID NO: 442)]-x-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)] [N is the active site residue]

[ 1] Dufour E. Biochimie 70:1335-1342(1988).

[ 2] Kirschke H., Barrett A.J., Rawlings N.D. Protein Prof. 2:1587-1643(1995).

[ 3] Shi G.-P., Chapman H.A., Bhairi S.M., Deleeuw C., Reddy V.Y., Weiss S.J. FEBS Lett. 357:129-134(1995).

[ 4] Velasco G., Ferrando A.A., Puente X.S., Sanchez L.M., Lopez-Otin C. J. Biol. Chem. 269:27136-27142(1994).

[ 5] Chapot-Chartier M.P., Nardi M., Chopin M.C., Chopin A., Gripon J.C. Appl. Environ. Microbiol. 59:330-333(1993).

[ 6] Higgins D.G., McConnell D.J., Sharp P.M. Nature 340:604-604(1989).

[ 7] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

104. Cys_Met_Meta_PP (Cys/Met metabolism PLP-dependent enzyme)

[1] Medline: 96428687

Crystal structure of the pyridoxal-5'-phosphate dependent cystathionine beta-lyase from Escherichia coli at 1.83 A.

Clausen T, Huber R, Laber B, Pohlenz HD, Messerschmidt A;

J Mol Biol 1996;262:202-224.

[1] Medline: 99059720

Crystal structure of Escherichia coli cystathionine gamma-synthase at 1.5 A resolution.

Clausen T, Huber R, Prade L, Wahl MC, Messerschmidt A;

EMBO J 1998;17:6827-6838.

Database Reference: SCOP; 1cs1; fa; [SCOP-USA][CATH-PDBSUM]

This family includes enzymes involved in cysteine and methionine metabolism. The following are members:

Cystathionine gamma-lyase,

Cystathionine gamma-synthase,

Cystathionine beta-lyase,

Methionine gamma-lyase,

OAH/OAS sulfhydrylase,

O-succinylhomoserine sulphhydrylase

All of these members participate is slightly different reactions.

All these enzymes use PLP (pyridoxal-5'-phosphate) as a cofactor.

Number of members: 52

A number of pyridoxal-dependent enzymes involved in the metabolism of cysteine, homocysteine and methionine have been shown [1,2] to be evolutionary related. These are:

- Cystathionine gamma-lyase (EC 4.4.1.1) (gamma-cystathionase), which catalyzes the transformation of cystathionine into cysteine, oxobutanoate and ammonia. This is the final reaction in the transulfuration pathway that leads from methionine to cysteine in eukaryotes.
- Cystathionine gamma-synthase (EC 4.2.99.9), which catalyzes the conversion of cysteine and succinyl-homoserine into cystathionine and succinate: the first step in the biosynthesis of methionine from cysteine in bacteria (gene metB).
- Cystathionine beta-lyase (EC 4.4.1.8) (beta-cystathionase), which catalyzes the conversion of cystathionine into homocysteine, pyruvate and ammonia: the second step in the biosynthesis of methionine from cysteine in bacteria (gene metC).
- Methionine gamma-lyase (EC 4.4.1.11) (L-methioninase) which catalyzes the transformation of methionine into methanethiol, oxobutanoate and ammonia.
- OAH/OAS sulfhydrylase, which catalyzes the conversion of acetylhomoserine into homocysteine and that of acetylserine into cysteine (gene MET17 or MET25 in yeast).
- O-succinylhomoserine sulfhydrylase (EC 4.2.99.-).
- Yeast hypothetical protein YGL184c.
- Yeast hypothetical protein YHR112c.

These enzymes are proteins of about 400 amino-acid residues. The pyridoxal-P group is attached to a lysine residue located in the central section of these enzymes; the sequence around this residue is highly conserved and can be used as a signature pattern to detect this class of enzymes.

-Consensus pattern: [DQ]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x(3)-~~[STAGC]~~[STAGC (SEQ ID NO: 691)]-~~[STAGCI]~~[STAGCI (SEQ ID NO: 693)]-T-K-~~[FYWQ]~~[FYWQ (SEQ ID NO: 159)]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x-G-[HQ]-~~[SGNH]~~[SGNH (SEQ ID NO: 679)] [K is the pyridoxal-P attachment site]

[1] Ono B.I., Tanaka K., Naito K., Heike C., Shinoda S., Yamamoto S., Ohmori S., Oshima T., Toh-E A.
J. Bacteriol. 174:3339-3347(1992).

[ 2] Barton A.B., Kaback D.B., Clark M.W., Keng T., Ouellette B.F.F., Storms R.K., Zeng B., Zhong W.W., Fortin N., Delaney S., Bussey H. Yeast 9:363-369(1993).

105. Cyt_reductase

FAD/NAD-binding Cytochrome reductase

Number of members: 60

[1] Medline: 95111952

Crystal structure of the FAD-containing fragment of corn nitrate reductase at 2.5 A resolution: relationship to other flavoprotein reductases.

Lu G, Campbell WH, Schneider G, Lindqvist Y;

Structure 1994;2:809-821.

[2] Medline: 92084635

The sequence of squash NADH:nitrate reductase and its relationship to the sequences of other flavoprotein oxidoreductases. A family of flavoprotein pyridine nucleotide cytochrome reductases.

Hyde GE, Crawford NM, Campbell WH;

J Biol Chem 1991;266:23542-23547.

106. Cytidylyltrans

Phosphatidate cytidylyltransferase

Number of members: 21

Phosphatidate cytidylyltransferase (EC 2.7.7.41) [1,2,3] (also known as CDP-diacylglycerol synthase) (CDS) is the enzyme that catalyzes the synthesis of CDP-diacylglycerol from CTP and phosphatidate (PA). CDP-diacylglycerol is an important branch point intermediate in both prokaryotic and eukaryotic organisms. CDS is a membrane-bound enzyme. A conserved region located in the C-terminal part has been selected as a signature pattern.

-Consensus pattern: S-x-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-K-R-x(4)-K-D-x-[GSA]-x(2)-[LI]-[PG]-x-H-G-G-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-D-R-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-D

[ 1] Sparrow C.P., Raetz C.R.H.
J. Biol. Chem. 260:12084-12091(1985).
[ 2] Shen H., Heacock P.N., Clancey C.J., Dowhan W.
J. Biol. Chem. 271:789-795(1996).
[ 3] Saito S., Goto K., Tonosaki A., Kondo H.
J. Biol. Chem. 272:9503-9509(1997).

107. (Cytidylyltransf) Cytidylyltransferase. This family includes: Cholinephosphate cytidylyltransferase. Glycerol-3-phosphate cytidylyltransferase.
Number of members: 64

[1] Medline: 10208837 CTP:Phosphocholine Cytidylyltransferase: Insights into Regulatory Mechanisms and Novel Functions. Clement JM, Kent C; Biochem Biophys Res Commun 1999;257:643-650.

108. (cNMP binding) Cyclic nucleotide-binding domain signatures and profile
Proteins that bind cyclic nucleotides (cAMP or cGMP) share a structural domainof about 120 residues [1-3]. The best studied of these proteins is theprokaryotic catabolite gene activator (also known as the cAMP receptorprotein) (gene crp) where such a domain is known to be composed of threealpha-helices and a distinctive eight-stranded, antiparallel beta-barrelstructure. Such a domain is known to exist in the following proteins: - Prokaryotic catabolite gene activator protein (CAP). - cAMP- and cGMP-dependent protein kinases (cAPK and cGPK). Both types of kinases contains two tandem copies of the cyclic nucleotide-binding domain. The cAPK's are composed of two different subunits: a catalytic chain and a regulatory chain which contains both copies of the domain. The cGPK's are single chain enzymes that include the two copies of the domain in their N- terminal section. The nucleotide specificity of cAPK and cGPK is due to an amino acid in the conserved region of beta-barrel 7: a threonine that is invariant in cGPK is an alanine in most cAPK. -

Vertebrate cyclic nucleotide-gated ion-channels. Two such cations channels have been fully characterized. One is found in rod cells where it plays a role in visual signal transduction. It specifically binds to cGMP leading to an opening of the channel and thereby causing a depolarization of rod photoreceptors. In olfactory epithelium a similar, cAMP-binding, channel plays a role in odorant signal transduction. There are six invariant amino acids in this domain, three of which are glycine residues that are thought to be essential for maintenance of the of the beta-barrel. Two signature patterns for this domain have been developed. The first pattern is located within beta-barrels 2 and 3 and contains the first two conserved Gly. The second pattern is located within beta-barrels 6 and 7 and contains the third conserved Gly as well as the three other invariant residues.-

First consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[VIC]-x(2)-G-~~[DENQTA]~~[DENQTA (SEQ ID NO: 55)]-x-[GAC]-x(2)-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)](4)- x(2)-G

Second consensus pattern: ~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-G-E-x-[GAS]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(5,11)-R-~~[STAQ]~~[STAQ (SEQ ID NO: 730)]-A-x-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)]-x- ~~[STACV]~~[STACV (SEQ ID NO: 686)]-

[ 1] Weber I.T., Shabb J.B., Corbin J.D. Biochemistry 28:6122-6127(1989).

[ 2] Kaupp U.B. Trends Neurosci. 14:150-157(1991).

[ 3] Shabb J.B., Corbin J.D. J. Biol. Chem. 267:5723-5726(1992).

109. (cadherin)

Cadherins extracellular repeated domain signature

Cadherins [1,2] are a family of animal glycoproteins responsible for calcium-dependent cell-cell adhesion. Cadherins preferentially interact with themselves in a homophilic manner in connecting cells; thus acting as both receptor and ligand. A wide number of tissue-specific forms of cadherins are known:

- Epithelial (E-cadherin) (also known as uvomorulin or L-CAM) (CDH1).
- Neural (N-cadherin) (CDH2).
- Placental (P-cadherin) (CDH3).
- Retinal (R-cadherin) (CDH4).
- Vascular endothelial (VE-cadherin) (CDH5).

- Kidney (K-cadherin) (CDH6).
- Cadherin-8 (CDH8).
- Osteoblast (OB-cadherin) (CDH11).
- Brain (BR-cadherin) (CDH12).
- T-cadherin (truncated cadherin) (CDH13).
- Muscle (M-cadherin) (CDH14).
- Liver-intestine (LI-cadherin).
- EP-cadherin.

Structurally, cadherins are built of the following domains: a signal sequence, followed by a propeptide of about 130 residues, then an extracellular domain of around 600 residues, then a transmembrane region, and finally a C-terminal cytoplasmic domain of about 150 residues. The extracellular domain can be sub- divided into five parts: there are four repeats of about 110 residues followed by a region that contains four conserved cysteines. It is suggested that the calcium-binding region of cadherins is located in the extracellular repeats.

Cadherins are evolutionary related to the desmogleins which are component of intercellular desmosome junctions involved in the interaction of plaque proteins:

- Desmoglein 1 (desmosomal glycoprotein I).
- Desmoglein 2.
- Desmoglein 3 (Pemphigus vulgaris antigen).

The Drosophila fat protein [3] is a huge protein of over 5000 amino acids that contains 34 cadherin-like repeats in its extracellular domain.

The signature pattern that was developed for the repeated domain is located in it the C-terminal extremity which is its best conserved region. The pattern includes two conserved aspartic acid residues as well as two asparagines; these residues could be implicated in the binding of calcium.

Consensus pattern[LIV]-x-[LIV]-x-D-x-N-D-[NH]-x-P Sequences known to belong to this class detected by the pattern ALL. Note this pattern is found in the first, second, and fourth

copies of the repeated domain. In the third copy there is a deletion of one residue after the second conserved Asp.

[ 1] Takeichi M. Annu. Rev. Biochem. 59:237-252(1990).

[ 2] Takeichi M. Trends Genet. 3:213-217(1987).

[ 3] Mahoney P.A., Weber U., Onofrechuk P., Biessmann H., Bryant P.J., Goodman C.S. Cell 67:853-868(1991).

110. Calreticulin family signatures

Calreticulin [1] (also known as calregulin, CRP55 or HACBP) is a high-capacitycalcium-binding protein which is present in most tissues and located at the periphery of the endoplasmic (ER) and the sarcoplamic reticulum (SR)membranes. It probably plays a role in the storage of calcium in the lumen ofthe ER and SR and it may well have other important functions. Structurally, calreticulin is a protein of about 400 amino acid residues consisting of three domains: a) An N-terminal, probably globular, domain of about 180 amino acid residues (N-domain); b) A central domain of about 70 residues (P-domain) which contains three repeats of an acidic 17 amino acid motif. This region binds calcium with a low-capacity, but a high-affinity; c) A C-terminal domain rich in acidic residues and in lysine (C-domain). This region binds calcium with a high-capacity but a low-affinity. Calreticulin is evolutionary related to the following proteins: - Onchocerca volvulus antigen RAL-1. RAL-1 is highly similar to calreticulin, but possesses a C-terminal domain rich in lysine and arginine and lacks acidic residues and is therefore not expected to bind calcium in that region. - Calnexin [2]. A calcium-binding protein that interacts with newly synthesized glycoproteins in the endoplasmic reticulum. It seems to play a major role in the quality control apparatus of the ER by the retention of incorrectly folded proteins. - Calmegin [3] (or calnexin-T), a testis-specific calcium-binding protein highly similar to calnexin. Three signature patterns have been developed for this family of proteins. The first two patterns are based on conserved regions in the N-domain; the third pattern corresponds to positions 4 to 16 of the repeated motif in the P-domain.

Consensus pattern: ~~[KRHN]~~[KRHN (SEQ ID NO: 300)]-x-~~[DEQN]~~[DEQN (SEQ ID NO: 74)]-~~[DEQNK]~~[DEQNK (SEQ ID NO: 76)]-x(3)-C-G-G-[AG]-[FY]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[KN]- ~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)](2)-

Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-F-G-P-D-x-C-[AG]-

Consensus pattern: [IV]-x-D-x-~~[DENST]~~[DENST (SEQ ID NO: 60)]-x(2)-K-P-[DEH]-D-W-[DEN]-

[ 1] Michalak M., Milner R.E., Burns K., Opas M. Biochem. J. 285:681-692(1992).

[ 2] Bergeron J.J.M., Brenner M.B., Thomas D.Y., Williams D.B. Trends Biochem. Sci. 19:124-128(1994).

[ 3] Watanabe D., Yamada K., Nishina Y., Tajima Y., Koshimizu U., Nagata A., Nishimune Y. J. Biol. Chem. 269:7744-7749(1994).

111. Eukaryotic-type carbonic anhydrases signature (carb_anhydrase)

Carbonic anhydrases (EC 4.2.1.1) (CA) [1,2,3,4] are zinc metalloenzymes which catalyze the reversible hydration of carbon dioxide. Eight enzymatic and evolutionary related forms of carbonic anhydrase are currently known to exist in vertebrates: three cytosolic isozymes (CA-I, CA-II and CA-III); two membrane-bound forms (CA-IV and CA-VII); a mitochondrial form (CA-V); a secreted salivary form (CA-VI); and a yet uncharacterized isozyme [5].In the alga Chlamydomonas reinhardtii, two CA isozymes have been sequenced[6]. They are periplasmic glycoproteins evolutionary related to vertebrate CAs. Some bacteria, such as Neisseria gonorrhoeae [7] also have a eukaryotic-type CA.CAs contain a single zinc atom bound to three conserved histidine residues. As a signature for CAs, a pattern has been developed which includes one of these zinc-binding histidines. Protein D8 from Vaccinia and other poxviruses is related to CAs but has lost two of the zinc-binding histidines as well as many otherwise conserved residues. This is also true of the N-terminal extracellular domain of some receptor-type tyrosine-protein phosphatases (see <PDOC00323>).

Consensus pattern: S-E-[HN]-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(4)-[FYH]-x(2)-E-~~[LIVMGA]~~[LIVMGA (SEQ ID NO: 487)]-H-~~[LIVMFA]~~[LIVMFA (SEQ ID NO: 403)](2) [The second H is a zinc ligand]-

Note: most prokaryotic CA's as well as plant chloroplast CA's belong to another, evolutionary distinct family of proteins (see <PDOC00586

[ 1] Deutsch H.F. Int. J. Biochem. 19:101-113(1987).

[ 2] Fernley R.T. Trends Biochem. Sci. 13:356-359(1988).

[ 3] Tashian R.E. BioEssays 10:186-192(1989).

[ 4] Edwards Y. Biochem. Soc. Trans. 18:171-175(1990).

[ 5] Skaggs L.A., Bergenhem N.C.H., Venta P.J., Tashian R.E. Gene 126:291-292(1993).

[ 6] Fujiwara S., Fukuzawa H., Tachiki A., Miyachi S. Proc. Natl. Acad. Sci. U.S.A. 87:9779-9783(1990).

[ 7] Huang S., Xue Y., Sauer-Eriksson E., Chirica L., Lindskog S.,Jonsson B.H. 2.3.CO;2-"J. Mol. Biol. 283:301-310(1998).

112. Caseins alpha/beta signature

Caseins [1] are the major protein constituent of milk. Caseins can be classified into two families; the first consists of the kappa-caseins, and the second groups the alpha-s1, alpha-s2, and beta-caseins. The alpha/beta caseins are a rapidly diverging family of proteins. However two regions are conserved: a cluster of phosphorylated serine residues and the signal sequence. The signature pattern has been developed for this family of proteins based upon the last eight residues of the signal sequence.

Consensus pattern: C-L-[LV]-A-x-A-[LVF]-A –

[ 1] Holt C., Sawyer L. Protein Eng. 2:251-259(1988).

113. Catalase signatures

Catalase (EC 1.11.1.6) [1,2,3] is an enzyme, present in all aerobic cells,that decomposes hydrogen peroxide to molecular oxygen and water. Its main function is to protect cells from the toxic effects of hydrogen peroxide. In eukaryotic organisms and in some prokaryotes catalase is a molecule composed of four identical subunits. Each of the subunits binds one protoheme IX group. A conserved tyrosine serves as the heme proximal side ligand. The region around this residue has been used as a first signature pattern; it also includes a conserved arginine that participates in heme-binding. A conserved histidine has been shown to be important for the catalytic mechanism of the enzyme. The region around this residue has been selected as a second signature pattern.-

Consensus pattern: R-~~[LIVMFSTAN]~~[LIVMFSTAN (SEQ ID NO: 424)]-F-~~[GASTNP]~~[GASTNP (SEQ ID NO: 177)]-Y-x-D-[AST]-[QEH] [Y is the proximal heme-binding ligand]

Consensus pattern: [IF]-x-[RH]-x(4)-[EQ]-R-x(2)-H-x(2)-[GAS]-~~[GASTF]~~[GASTF (SEQ ID NO: 176)]-~~[GAST]~~[GAST (SEQ ID NO: 175)] [H is an active site residue]

Note: some prokaryotic catalases belong to the peroxidase family (see <PDOC00394>).

[ 1] Murthy M.R.N., Reid T.J. III, Sicignano A., Tanaka N., Rossmann M.G. J. Mol. Biol. 152:465-499(1981).

[ 2] Melik-Adamyan W.R., Barynin V.V., Vagin A.A., Borisov V.V., Vainshtein B.K., Fita I., Murthy M.R.N., Rossmann M.G. J. Mol. Biol. 188:63-72(1986).

[ 3] von Ossowki I., Hausner G., Loewen P.C. J. Mol. Evol. 37:71-76(1993).

114. (chitin binding) Chitin recognition or binding domain signature

A conserved domain of 43 amino acids is found in several plant and fungal proteins that have a common binding specificity for oligosaccharides of N-acetylglucosamine [1]. This domain may be involved in the recognition or binding of chitin subunits. It has been found in the proteins listed below. - A number of non-leguminous plant lectins. The best characterized of these lectins are the three highly homologous wheat germ agglutinins (WGA-1, 2 and 3). WGA is an N-acetylglucosamine/N-acetylneuraminic acid binding lectin which structurally consists of a fourfold repetition of the 43 amino acid domain. The same type of structure is found in a barley root-specific lectin as well as a rice lectin. - Plants endochitinases (EC 3.2.1.14) from class IA (see <PDOC00620>). Endochitinases are enzymes that catalyze the hydrolysis of the beta-1,4 linkages of N-acetyl glucosamine polymers of chitin. Plant chitinases function as a defense against chitin containing fungal pathogens. Class IA chitinases generally contain one copy of the chitin-binding domain at their N-terminal extremity. An exception is agglutinin/chitinase [2] from the stinging nettle Urtica dioica which contains two copies of the domain. - Hevein [5], a wound-induced protein found in the latex of rubber trees. - Win1 and win2, two wound-induced proteins from potato. - Kluyveromyces lactis killer toxin alpha subunit [3]. The toxin encoded by the linear plasmid pGKL1 is composed of three subunits: alpha, beta, and gamma. The gamma subunit harbors toxin activity and inhibits growth of sensitive yeast strains in the G1 phase of the cell cycle; the alpha subunit, which is proteolytically processed from a larger precursor that also contains the beta subunit, is a chitinase (see <PDOC00839>). In chitinases, as well as in the potato wound-induced proteins, the 43-residuedomain directly follows the signal sequence and is therefore at the N-terminal of the mature protein; in the killer toxin alpha subunit it is located in the central section of the protein. The domain contains eight conserved cysteine residues which have all been shown, in WGA, to be involved in disulfide bonds. The

topological arrangement of the four disulfide bonds is shown in the following figure: +--------
-----+ +----|------+ | | | | | xxCgxxxxxxxCxxxxCCsxxgxCgxxxxxCxxxCxxxxC |
******|********** | | | | +----+ +----------+'C': conserved cysteine involved in a
disulfide bond.'*': position of the pattern.

-Consensus pattern: C-x(4,5)-C-C-S-x(2)-G-x-C-G-x(4)-[FYW]-C [The five C's are involved in disulfide bonds]

[ 1] Wright H.T., Sandrasegaram G., Wright C.S. J. Mol. Evol. 33:283-294(1991).
[ 2] Lerner D.R., Raikhel N.V. J. Biol. Chem. 267:11085-11091(1992).
[ 3] Butler A.R., O'Donnel R.W., Martin V.J., Gooday G.W., Stark M.J.R. Eur. J. Biochem. 199:483-488(1991).

115. (Chitinase 1) Chitinases family 19 signatures
Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of thebeta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the viewpoint of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 19(also known as classes IA or I and IB or II) are enzymes from plants that function in the defense against fungal and insect pathogens by destroying their chitin-containing cell wall. Class IA/I and IB/II enzymes differ in the presence (IA/I) or absence (IB/II) of a N-terminal chitin-binding domain (seethe relevant entry <PDOC00025>). The catalytic domain of these enzymes consist of about 220 to 230 amino acid residues. Two highly conserved regions have been selected as signature patterns, the first one is located in the N-terminal section and contains one of the six cysteines which are conserved in most, if not all, of these chitinases and which is probably involved in a disulfide bond.

Consensus pattern: C-x(4,5)-F-Y-[ST]-x(3)-[FY]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x-A-x(3)-[YF]-x(2)-F- [GSA]

Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[GSA]-F-x-~~[STAG]~~[STAG (SEQ ID NO: 690)](2)-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-W-[FY]-W-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]

[ 1] Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).

[ 2] Henrissat B. Biochem. J. 280:309-316(1991).

116. chloroa_b-bind

Chlorophyll A-B binding proteins. Number of members: 211

117. chromo

The 'chromo' (CHRromatin Organization MOdifier) domain [1 to 4] is a conserved region of about 60 amino acids which was originally found in Drosophila modifiers of variegation, which are proteins that modify the structure of chromatin to the condensed morphology of heterochromatin, a cytologically visible condition where gene expression is repressed. In protein Polycomb, the chromo domain has been shown to be important for chromatin targeting. Proteins that contains a chromo domain seem to fall into three classes:

a) Proteins which have a N-terminal chromo domain followed by a region which is related to but distinct from the chromo domain and which has been termed [3] the 'chromo shadow' domain.
b) Proteins with a single chromo domain.
c) Proteins with paired tandem chromo domains.

Currently, this domain has been found in the following proteins:

Class A.

- Drosophila heterochromatin protein Su(var)205 (HP1).
- Human heterochromatin protein HP1 alpha.
- Mammalian modifier 1 and modifier 2.
- Fission yeast swi6, a protein involved in the repression of the silent mating-type loci mat2 and mat3.

Class B.

- Drosophila protein Polycomb (Pc).
- Mammalian modifier 3, a homolog of Pc.

- Drosophila protein Su(var)3-9, a suppressor of position-effect variegation.
- Human Mi-2 autoantigen, characterisitic of dermatomyosis.
- Fungal retrotranposon polyproteins: 'skippy' from Fusarium oxysporum, 'grasshopper' and 'MAGGY' from Magnaporthe grisea and CfT-1 from Cladosporium fulvum.
- Fission yeast hypothetical protein SpAC18G6.02c.
- Caenorhabditis elegans hypothetical protein C29H12.5
- Caenorhabditis elegans hypothetical protein ZK1236.2.
- Caenorhabditis elegans hypothetical protein T09A5.8.

Class C.

- Mammalian DNA-binding/helicase proteins CHD-1 to CHD-4.
- Yeast protein CHD1.

The signature pattern for this domain corresponds to its best conserved section, which is located in its central part.

-Consensus pattern: [FYL]-x-~~[LIVMC]~~[LIVMC (SEQ ID NO: 396)]-[KR]-W-x-~~[GDNR]~~[GDNR (SEQ ID NO: 182)]-~~[FYWLME]~~[FYWLME (SEQ ID NO: 152)]-x(5,6)-[ST]-W-[ESV]-~~[PSTDEN]~~[PSTDEN (SEQ ID NO: 619)]-x(2,3)-~~[LIVMC]~~[LIVMC (SEQ ID NO: 396)]

[ 1] Paro R. Trends Genet. 6:416-421(1990).

[ 2] Singh P.B., Miller J.R., Pearce J., Kothary R., Burton R.D., Paro R., James T.C., Gaunt S.J. Nucleic Acids Res. 19:789-794(1991).

[ 3] Aasland R., Stewart A.F. Nucleic Acids Res. 23:3168-3173(1995).

[ 4] Koonin E.V., Zhou S., Lucchesis J.C. Nucleic Acids Res. 23:4229-4233(1995).

118. citrate_synt

Citrate synthase (EC 4.1.3.7) (CS) is the tricarboxylic acid cycle enzyme that catalyzes the synthesis of citrate from oxaloacetate and acetyl-CoA in an aldol condensation. CS can directly form a carbon-carbon bond in the absence of metal ion cofactors.

In prokaryotes, citrate synthase is composed of six identical subunits. In eukaryotes, there are two isozymes of citrate synthase: one is found in the mitochondrial matrix, the second is cytoplasmic. Both seem to be dimers of identical chains.

There are a number of regions of sequence similarity between prokaryotic and eukaryotic citrate synthases. One of the best conserved contains a histidine which is one of three residues shown [1] to be involved in the catalytic mechanism of the vertebrate mitochondrial enzyme. This region has been used as a signature pattern.

-Consensus pattern: G-[FYA]-[GA]-H-x-[IV]-x(1,2)-[RKT]-x(2)-D-[PS]-R [H is an active site residue]

[ 1] Karpusas M., Branchaud B., Remington S.J. Biochemistry 29:2213-2219(1990).

119. clpA_B

Chaperonin clpA/B

CAUTION! This family is a subfamily of the AAA superfamily. The threshold has been set very high to stop overlaps with the AAA superfamily. This entry will be subsumed by AAA in the future.

Number of members: 39

A number of ATP-binding proteins that are are thought to protect cells from extreme stress by controlling the aggregation of denaturation of vital cellular structures have been shown [1,2] to be evolutionary related. These proteins are listed below.

- Escherichia coli clpA, which acts as the regulatory subunit of the ATP-dependent protease clp.
- Rhodopseudomonas blastica clpA homolog.

- Escherichia coli heat shock protein clpB and homologs in other bacteria.
- Bacillus subtilis protein mecB.
- Yeast heat shock protein 104 (gene HSP104), which is vital for tolerance to heat, ethanol and other stresses.
- Neurospora heat shock protein hsp98.
- Yeast mitochondrial heat shock protein 78 (gene HSP78) [3].
- CD4A and CD4b, two highly related tomato proteins that seem to be located in the chloroplast.
- Trypanosoma brucei protein clp.
- Porphyra purpurea chloroplast encoded clpC.

The size of these proteins range from 84 Kd (clpA) to slightly more than 100 Kd (HSP104). They all share two conserved regions of about 200 amino acids that each contains an ATP-binding site. In addition to the ATP-binding A and B motifs there are many parts in these two domains that are also conserved. Two of these regions have been selected as signature patterns. The first signature is located in the first domain, some ten residues to the C-terminal of the ATP-binding B motif. The second pattern is located in the second domain in-between the ATP-binding A and B motifs.

-Consensus pattern: D-[AI]-[SGA]-N-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)](2)-K-[PT]-x-L-x(2)-G

-Consensus pattern: R-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-D-x-S-E-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x-E-[KRQ]-x-[STA]-x-[STA]-[KR]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-G-[STA]

[ 1] Gottesman S., Squires C., Pichersky E., Carrington M., Hobbs M., Mattick J.S., Dalrymple B., Kuramitsu H., Shiroza T., Foster T., Clark W.P., Ross B., Squires C.L., Maurizi M.R. Proc. Natl. Acad. Sci. U.S.A. 87:3513-3517(1990).

[ 2] Parsell D.A., Sanchez Y., Stitzel J.D., Lindquist S. Nature 353:270-273(1991).

[ 3] Leonhardt S.A., Fearon K., Danese P.N., Mason T.L. Mol. Cell. Biol. 13:6304-6313(1993).

120. cofilin_ADF

Cofilin/tropomyosin-type actin-binding proteins

[1]

Medline: 97290449

Structure determination of yeast cofilin.

Fedorov AA, Lappalainen P, Fedorov EV, Drubin DG, Almo SC;

Nat Struct Biol 1997;4:366-369.

[2]

Medline: 97290450

Crystal structure of the actin-binding protein actophorin from Acanthamoeba.

Leonard SA, Gittis AG, Petrella EC, Pollard TD, Lattman EE;

Nat Struct Biol 1997;4:369-373.

[3]

Medline: 97420794

F-actin and G-actin binding are uncoupled by mutation of conserved tyrosine residues in maize actin depolymerizing factor.

Jiang CJ, Weeds AG, Khan S, Hussey PJ;

Proc Natl Acad Sci U S A 1997;94:9973-9978.

[4]

Medline: 97357155

Cofilin promotes rapid actin filament turnover in vivo.

Lappalainen P, Drubin DG;

Nature 1997;388:78-82.

Severs actin filaments and binds to actin monomers.

Number of members: 44

Actin-depolymerizing proteins sever actin filaments (F-actin) and/or bind to actin monomers, or G-actin, thus preventing actin-polymerization by sequestering the monomers. The following proteins are evolutionary related and belong to a family of low molecular weight (137 to 166 residues) actin-depolymerizing proteins [1,2,3,4]:

- Cofilin from vertebrates, slime mold and yeast. Cofilin binds to F-actin and acts as a pH-dependent actin-depolymerizing protein.
- Destrin from vertebrates. Destrin binds to G-actin in a pH-independent manner and prevents polymerization.
- Caenorhabditis elegans unc-60.
- Acanthamoeba castellanii actophorin.
- Plants actin depolymerizing factor (ADF).

The most conserved region of these proteins is a twenty amino-acid segment that ends some 30 residues from their C-terminal extremity. This segment has been shown [5] to be important for actin-binding.

-Consensus pattern: P-[DE]-x-[SA]-x-~~[LIVMT]~~[LIVMT (SEQ ID NO: 518)]-[KR]-x-[KR]-M-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[YA]-[STA](3)-x(3)-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-[KR]

[ 1] Hawkins M., Pope B., MacIver S.K., Weeds A.G. Biochemistry 32:9985-9993(1993).
[ 2] Iida K., Moriyama K., Matsumoto S., Kawasaki H., Nishida E., Yahara I. Gene 124:115-120(1993).
[ 3] Quirk S., MacIver S.K., Ampe C., Doberstein S.K., Kaiser D.A., van Damme J., Vandekerckhove J., Pollard T.D. Biochemistry 32:8525-8533(1993).
[ 4] McKim K.S., Matheson C., Marra M.A., Wakarchuk M.F., Baillie D.L. Mol. Gen. Genet. 242:346-357(1994).
[ 5] Moriyama K., Yonezawa N., Sakai H., Yahara I., Nishida E. J. Biol. Chem. 267:7240-7244(1992).

121. (Complex 24kd) Respiratory-chain NADH dehydrogenase 24 Kd subunit signature
Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complexI or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist inthe chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase).Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 24 Kd (in mammals), which is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind a2Fe-

2S iron-sulfur cluster. The 24 Kd subunit is nuclear encoded, as aprecursor form with a transit peptide in mammals, and in Neurospora crassa.The 24 Kd subunit is highly similar to [3,4]: - Subunit E of Escherichia coli NADH-ubiquinone oxidoreductase (gene nuoE). - Subunit NQO2 of Paracoccus denitrificans NADH-ubiquinone oxidoreductase. A highly conserved region, located in the central section of this subunit containing two conserved cysteines that are probably involved in the binding of the 2Fe-2S center has been selected as a signature pattern.

-Consensus pattern: D-x(2)-F-[ST]-x(5)-C-L-G-x-C-x(2) [GA]-P [The two C's are putative 2Fe-2S ligands]

[ 1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[ 2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

[ 3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).

[ 4] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

122. copper-bind

Copper binding proteins, plastocyanin/azurin family

Number of members: 70

Blue or 'type-1' copper proteins are small proteins which bind a single copper atom and which are characterized by an intense electronic absorption band near 600 nm [1,2]. The most well known members of this class of proteins are the plant chloroplastic plastocyanins, which exchange electrons with cytochrome c6, and the distantly related bacterial azurins, which exchange electrons with cytochrome c551. This family of proteins also includes all the proteins listed below (references are only provided for recently determined sequences).

- Amicyanin from bacteria such as Methylobacterium extorquens or Thiobacillus versutus that can grow on methylamine. Amicyanin appears to be an electron receptor for methylamine dehydrogenase.
- Auracyanins A and B from Chloroflexus aurantiacus [3]. These proteins can

donate electrons to cytochrome c-554.

- Blue copper protein from Alcaligenes faecalis.
- Cupredoxin (CPC) from cucumber peelings [4].
- Cusacyanin (basic blue protein; plantacyanin, CBP) from cucumber.
- Halocyanin from Natrobacterium pharaonis [5], a membrane associated copper-binding protein.
- Pseudoazurin from Pseudomonas.
- Rusticyanin from Thiobacillus ferrooxidans. Rusticyanin is an electron carrier from cytochrome c-552 to the a-type oxidase [6].
- Stellacyanin from the Japanese lacquer tree.
- Umecyanin from horseradish roots.

- Allergen Ra3 from ragweed. This pollen protein is evolutionary related to the above proteins, but seems to have lost the ability to bind copper.

Although there is an appreciable amount of divergence in the sequence of all these proteins, the copper ligand sites are conserved and a pattern which includes two of the ligands (a cysteine and a histidine) has been developed.

-Consensus pattern: [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ] [C and H are copper ligands]

[ 1] Garret T.P.J., Clingeleffer D.J., Guss J.M., Rogers S.J., Freeman H.C. J. Biol. Chem. 259:2822-2825(1984).

[ 2] Ryden L.G., Hunt L.T. J. Mol. Evol. 36:41-66(1993).

[ 3] McManus J.D., Brune D.C., Han J., Sanders-Loehr J., Meyer T.E., Cusanovich M.A., Tollin G., Blankenship R.E. J. Biol. Chem. 267:6531-6540(1992).

[ 4] Mann K., Schaefer W., Thoenes U., Messerschmidt A., Mehrabian Z., Nalbandyan R. FEBS Lett. 314:220-223(1992).

[ 5] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

[ 6] Yano T., Fukumori Y., Yamanaka T. FEBS Lett. 288:159-162(1991).

123. Chaperonins cpn10 signature

Chaperonins [1,2] are proteins involved in the folding of proteins or the assembly of oligomeric protein complexes. They seem to assist other polypeptides in maintaining or assuming conformations which permit their correct assembly into oligomeric structures. They are found in abundance in prokaryotes, chloroplasts and mitochondria. Chaperonins form oligomeric complexes and are composed of two different types of subunits: a 60 Kd protein, known as cpn60 (groEL in bacteria) and a 10 Kd protein, known ascpn10 (groES in bacteria).The cpn10 protein binds to cpn60 in the presence of MgATP and suppresses the ATPase activity of the latter. Cpn10 is a protein of about 100 amino acid residues whose sequence is well conserved in bacteria, vertebrate mitochondriaand plants chloroplast [3,4]. Cpn10 assembles as an heptamer that forms a dome[5]. As a signature pattern for cpn10 , a region located in the N-terminal section of the protein was selected.

Consensus pattern: ~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x-P-[ILT]-x-[DEN]-[KR]-~~[LIVMFA]~~[LIVMFA (SEQ ID NO: 403)](3)-~~[KREQ]~~[KREQ (SEQ ID NO: 292)]-x(8,9)-[SG]-x-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)](3)-

Note: this pattern is found twice in the plant chloroplast protein which consist of the tandem repeat of a cpn10 domain

[ 1] Ellis R.J., van der Vies S.M. Annu. Rev. Biochem. 60:321-347(1991).

[ 2] Zeilsta-Ryalls J., Fayet O., Georgopoulos C. Annu. Rev. Microbiol. 45:301-325(1991).

[ 3] Hartman D.J., Hoogenraad N.J., Condron R., Hoj P.B. Proc. Natl. Acad. Sci. U.S.A. 89:3394-3398(1992).

[ 4] Bertsch U., Soll J., Seetharam R., Viitanen P.V. Proc. Natl. Acad. Sci. U.S.A. 89:8696-8700(1992).

[ 5] Hunt J.F., Weaver A.J., Landry S.J., Gierasch L., Deisenhofer J. Nature 379:37-45(1996).

124. Chaperonins cpn60 signature (cpn60_TCP1)

Chaperonins [1,2] are proteins involved in the folding of proteins or the assembly of oligomeric protein complexes. Their role seems to be to assist other polypeptides to maintain or assume conformations which permit their correct assembly into oligomeric structures. They are found in abundance in prokaryotes, chloroplasts and mitochondria. Chaperonins form oligomeric complexes and are composed of two different types of subunits: a 60 Kd

protein, known as cpn60 (groEL in bacteria) and a 10 Kd protein, known as cpn10 (groES in bacteria).The cpn60 protein shows weak ATPase activity and is a highly conserved protein of about 550 to 580 amino acid residues which has been described by different names in different species: - Escherichia coli groEL protein, which is essential for the growth of the bacteria and the assembly of several bacteriophages. - Cyanobacterial groEL analogues. - Mycobacterium tuberculosis and leprae 65 Kd antigen, Coxiella burnetti heat shock protein B (gene htpB), Rickettsia tsutsugamushi major antigen 58, and Chlamydial 57 Kd hypersensitivity antigen (gene hypB). - Chloroplast RuBisCO subunit binding-protein alpha and beta chains, which bind ribulose bisphosphate carboxylase small and large subunits and are implicated in the assembly of the enzyme oligomer. - Mammalian mitochondrial matrix protein P1 (mitonin or P60). - Yeast HSP60 protein, a mitochondrial assembly factor. As a signature pattern for these proteins, a rather well-conserved region of twelve residues, located in the last third of the cpn60sequence was chosen.

Consensus pattern: A-[AS]-x-[DEQ]-E-x(4)-G-G-[GA]-

[ 1] Ellis R.J., van der Vies S.M. Annu. Rev. Biochem. 60:321-347(1991).
[ 2] Zeilsta-Ryalls J., Fayet O., Georgopoulos C. Annu. Rev. Microbiol. 45:301-325(1991).

Chaperonins TCP-1 signatures (cpn60_TCP1)

The TCP-1 protein [1,2] (Tailless Complex Polypeptide 1) was first identified in mice where it is especially abundant in testis but present in all cell types. It has since been found and characterized in many other mammalian species, in Drosophila and in yeast. TCP-1 is a highly conserved protein of about 60 Kd (556 to 560 residues) which participates in a hetero-oligomeric900 Kd double-torus shaped particle [3] with 6 to 8 other different subunits. These subunits, the chaperonin containing TCP-1 (CCT) subunit beta, gamma,delta, epsilon, zeta and eta are evolutionary related to TCP-1 itself [4,5].The CCT is known to act as a molecular chaperone for tubulin, actin and probably some other proteins. The CCT subunits are highly related to archebacterial counterparts: - TF55 and TF56 [6], a molecular chaperone from Sulfolobus shibatae. TF55 has ATPase activity, is known to bind unfolded polypeptides and forms a oligomeric complex of two stacked nine-membered rings. - Thermosome [7], from Thermoplasma acidophilum. The thermosome is composed of two subunits (alpha and beta) and also seems to be a chaperone with ATPase activity. It forms an oligomeric complex of eight-membered rings. The TCP-1 family of proteins are weakly, but significantly [8], related

to thecpn60/groEL chaperonin family (see <PDOC00268>).As signature patterns of this family of chaperonins, three conserved regions located in the N-terminal domain were chosen.

Consensus pattern: ~~[RKEL]~~[RKEL (SEQ ID NO: 634)]-[ST]-x-~~[LMFY]~~[LMFY (SEQ ID NO: 544)]-G-P-x-[GSA]-x-x-K-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)](2)-

Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-[TS]-[NK]-D-[GA]-~~[AVNHK]~~[AVNHK (SEQ ID NO: 13)]-[TAV]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-x(2)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x-[SNH]-[PQH]-

Consensus pattern: Q-[DEK]-x-x-~~[LIVMGTA]~~[LIVMGTA (SEQ ID NO: 494)]-[GA]-D-G-T-

[ 1] Ellis J. Nature 358:191-192(1992).

[ 2] Nelson R.J., Craig E.A. Curr. Biol. 2:487-489(1992).

[ 3] Lewis V.A., Hynes G.M., Zheng D., Saibil H., Willison K.R. Nature 358:249-252(1992).

[ 4] Kubota H., Hynes G., Carne A., Ashworth A., Willison K.R. Curr. Biol. 4:89-99(1994)

[ 5] Kim S., Willison K.R., Horwich A.L. Trends Biochem. Sci. 20:543-548(1994).

[ 6] Trent J.D., Nimmesgern E., Wall J.S., Hartl F.U., Horwich A.L. Nature 354:490-493(1991).

[ 7] Waldmann T., Lupas A., Kellermann J., Peters J., Baumeister W. Biol. Chem. Hoppe-Seyler 376:119-126(1995).

[ 8] Hemmingsen S.M. Nature 357:650-650(1992).

125. cyclin (Cyclins)

The cyclins include an internal duplication, which is related to that found in TFIIB and the RB protein.

[1]

Medline: 94203808

Evidence for a protein domain superfamily shared by the cyclins, TFIIB and RB/p107.

Gibson TJ, Thompson JD, Blocker A, Kouzarides T;

Nucleic Acids Res 1994;22:946-952.

[2]

Medline: 96164440

The crystal structure of cyclin A

Brown NR, Noble MEM, Endicott JA, Garman EF, Wakatsuki S, Mitchell E, Rasmussen B, Hunt T, Johnson LN;

Structure. 1995;3:1235-1247.

Complex of cyclin and cyclin dependant kinase.

[3]

Medline: 96313126

Structural basis of cyclin-dependant kinase activation by phosphorylation.

Russo AA, Jeffrey PD, Pavletich NP;

Nat Struct Biol. 1996;3:696-700.

Cyclins regulate cyclin dependant kinases (CDKs).

The most divergent prosite members have been included. Swiss:P22674 the Uracil-DNA glycosylase 2 is the highest noise and may be related but has not been included.

Number of members: 189

Cyclins [1,2,3] are eukaryotic proteins which play an active role in controlling nuclear cell division cycles. Cyclins, together with the p34 (cdc2) or cdk2 kinases, form the Maturation Promoting Factor (MPF). There are two main groups of cyclins:

- G2/M cyclins, essential for the control of the cell cycle at the G2/M (mitosis) transition. G2/M cyclins accumulate steadily during G2 and are abruptly destroyed as cells exit from mitosis (at the end of the M-phase).
- G1/S cyclins, essential for the control of the cell cycle at the G1/S (start) transition.

In most species, there are multiple forms of G1 and G2 cyclins. For example, in vertebrates, there are two G2 cyclins, A and B, and at least three G1 cyclins, C, D, and E.

A cyclin homolog has also been found in herpesvirus saimiri [4].

The best conserved region is in the central part of the cyclins' sequences, known as the 'cyclin-box'. From this, a 32 residue pattern has been derived.

-Consensus pattern: R-x(2)-~~[LIVMSA]~~[LIVMSA (SEQ ID NO: 507)]-x(2)-~~[FYWS]~~[FYWS (SEQ ID NO: 160)]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-x(8)-~~[LIVMFC]~~[LIVMFC (SEQ ID NO: 412)]-x(4)-~~[LIVMFYA]~~[LIVMFYA (SEQ ID NO: 435)]-x(2)-~~[STAGC]~~[STAGC (SEQ ID NO: 691)]-~~[LIVMFYQ]~~[LIVMFYQ (SEQ ID NO: 451)]-x-~~[LIVMFYC]~~[LIVMFYC (SEQ ID NO: 439)]-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-D-[RKH]-~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)]

[ 1] Nurse P. Nature 344:503-508(1990).

[ 2] Norbury C., Nurse P. Curr. Biol. 1:23-24(1991).

[ 3] Lew D.J., Reed S.I. Trends Cell Biol. 2:77-81(1992).

[ 4] Nicholas J., Cameron K.R., Honess R.W. Nature 355:362-365(1992).

126. Cystatin domain

This is a very diverse family. Attempts to define separate subfamilies have failed. Typically, either the N-terminal or C-terminal end is very divergent. But splitting into two domains would make very short families. Cathelicidins are related to this family but have not been included. Number of members: 147

Inhibitors of cysteine proteases [1,2,3], which are found in the tissues and body fluids of animals, in the larva of the worm Onchocerca volvulus [4], as well as in plants, can be grouped into three distinct but related families:

- Type 1 cystatins (or stefins), molecules of about 100 amino acid residues with neither disulfide bonds nor carbohydrate groups.
- Type 2 cystatins, molecules of about 115 amino acid residues which contain one or two disulfide loops near their C-terminus.
- Kininogens, which are multifunctional plasma glycoproteins.

They are the precursor of the active peptide bradykinin and play a role in blood coagulation by helping to position optimally prekallikrein and factor XI next to factor XII. They are also inhibitors of cysteine proteases. Structurally, kininogens are made of three contiguous type-2 cystatin domains, followed by an additional domain (of variable length)

which contains the sequence of bradykinin. The first of the three cystatin domains seems to have lost its inhibitory activity.

In all these inhibitors, there is a conserved region of five residues which has been proposed to be important for the binding to the cysteine proteases. The consensus pattern starts one residue before this conserved region.

-Consensus pattern: ~~[GSTEQKRV]~~[GSTEQKRV (SEQ ID NO: 243)]-Q-~~[LIVT]~~[LIVT (SEQ ID NO: 538)]-[VAF]-~~[SAGQ]~~[SAGQ (SEQ ID NO: 663)]-G-x-~~[LIVMNK]~~[LIVMNK (SEQ ID NO: 498)]-x(2)-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x-~~[LIVMFYA]~~[LIVMFYA (SEQ ID NO: 435)]-~~[DENQKRHSIV]~~[DENQKRHSIV (SEQ ID NO: 41)]

[1] Barrett A.J. Trends Biochem. Sci. 12:193-196(1987).

[2] Rawlings N.D., Barrett A.J. J. Mol. Evol. 30:60-71(1990).

[3] Turk V., Bode W. FEBS Lett. 285:213-219(1991).

[4] Lustigman S., Brotman B., Huima T., Prince A.M. Mol. Biochem. Parasitol. 45:65-76(1991).

127. cytochrome_c (Cytochrome c)

The Pfam entry does not include all prosite members.
The cytochrome 556 and cytochrome c' families are not included.
Number of members: 259

In proteins belonging to cytochrome c family [1], the heme group is covalently attached by thioether bonds to two conserved cysteine residues. The consensus sequence for this site is Cys-X-X-Cys-His and the histidine residue is one of the two axial ligands of the heme iron. This arrangement is shared by all proteins known to belong to cytochrome c family, which presently includes cytochromes c, c', c1 to c6, c550 to c556, cc3/Hmc, cytochrome f and reaction center cytochrome c.

-Consensus pattern: C-~~{CPWHF}~~{CPWHF (SEQ ID NO: 772)}-~~{CPWR}~~{CPWR (SEQ ID NO: 773)}-C-H-~~{CFYW}~~{CFYW (SEQ ID NO: 770)}

[ 1] Mathews F.S. Prog. Biophys. Mol. Biol. 45:1-56(1985).

128. (DAGKa) Diacylglycerol kinase accessory domain (presumed)

Diacylglycerol (DAG) is a second messenger that acts as a protein kinase C activator. This domain is assumed to be an accessory domain: its function is unknown.

[1] Sakane F, Yamada K, Kanoh H, Yokoyama C, Tanabe T, Nature 1990;344:345-348.[2] Sakane F, Imai S, Kai M, Wada I, Kanoh H, J Biol Chem 1996;271:8394-8401.[3] Schaap D, de Widt J, van der Wal J, Vandekerckhove J, van, Damme J, Gussow D, Ploegh HL, van Blitterswijk WJ, van der, Bend RL, FEBS Lett 1990;275:151-158. [4] Kanoh H, Yamada K, Sakane F, Trends Biochem Sci 1990;15:47-50.

129. (DAGKc) Diacylglycerol kinase catalytic domain (presumed)

Diacylglycerol (DAG) is a second messenger that acts as a protein kinase C activator. The catalytic domain is assumed from the finding of bacterial homologues.

[1] Sakane F, Yamada K, Kanoh H, Yokoyama C, Tanabe T, Nature 1990;344:345-348. [2] Sakane F, Imai S, Kai M, Wada I, Kanoh H, J Biol Chem 1996;271:8394-8401. [3] Schaap D, de Widt J, van der Wal J, Vandekerckhove J, van, Damme J, Gussow D, Ploegh HL, van Blitterswijk WJ, van der, Bend RL, FEBS Lett 1990;275:151-158. [4] Kanoh H, Yamada K, Sakane F, Trends Biochem Sci 1990;15:47-50.

130. D-amino_acid_oxidases_signature(DAO)

D-amino acid oxidase (EC 1.4.3.3) (DAMOX or DAO) is an FAD flavoenzyme that catalyzes the oxidation of neutral and basic D-amino acids into their corresponding keto acids. DAOs have been characterized and sequenced in fungi and vertebrates where they are known to be located in the peroxisomes. D-aspartate oxidase (EC 1.4.3.1) (DASOX) [1] is an enzyme, structurally related to DAO, which catalyzes the same reaction but is active only toward dicarboxylic D-amino acids. In DAO, a conserved histidine has been shown [2] to be important for the enzyme's catalytic activity. The conserved region around this residue has been developed as a signature pattern for these enzymes.

Consensus pattern: ~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-H-[NHA]-Y-G-x-[GSA](2)-x-G-x(5)-G-x-A [H is a probable active site residue]o-

[ 1] Negri A., Ceciliani F., Tedeschi G., Simonic T., Ronchi S. J. Biol. Chem. 267:11865-11871(1992).

[ 2] Miyano M., Fukui K., Watanabe F., Takahashi S., Tada M., Kanashiro M., Miyake Y. J. Biochem. 109:171-177(1991).

131. DEAD and DEAH box families ATP-dependent helicases signatures

A number of eukaryotic and prokaryotic proteins have been characterized [1,2,3] on the basis of their structural similarity. They all seem to be involved in ATP-dependent, nucleic-acid unwinding. Proteins currently known to belong to this family are: - Initiation factor eIF-4A. Found in eukaryotes, this protein is a subunit of a high molecular weight complex involved in 5'cap recognition and the binding of mRNA to ribosomes. It is an ATP-dependent RNA-helicase. - PRP5 and PRP28. These yeast proteins are involved in various ATP-requiring steps of the pre-mRNA splicing process. - Pl10, a mouse protein expressed specifically during spermatogenesis. - An3, a Xenopus putative RNA helicase, closely related to Pl10. - SPP81/DED1 and DBP1, two yeast proteins probably involved in pre-mRNA splicing and related to Pl10. - Caenorhabditis elegans helicase glh-1. - MSS116, a yeast protein required for mitochondrial splicing. - SPB4, a yeast protein involved in the maturation of 25S ribosomal RNA. - p68, a human nuclear antigen. p68 has ATPase and DNA-helicase activities in vitro. It is involved in cell growth and division. - Rm62 (p62), a Drosophila putative RNA helicase related to p68. - DBP2, a yeast protein related to p68. - DHH1, a yeast protein. - DRS1, a yeast protein involved in ribosome assembly. - MAK5, a yeast protein involved in maintenance of dsRNA killer plasmid. - ROK1, a yeast protein. - ste13, a fission yeast protein. - Vasa, a Drosophila protein important for oocyte formation and specification of embryonic posterior structures. - Me31B, a Drosophila maternally expressed protein of unknown function. - dbpA, an Escherichia coli putative RNA helicase. - deaD, an Escherichia coli putative RNA helicase which can suppress a mutation in the rpsB gene for ribosomal protein S2. - rhlB, an Escherichia coli putative RNA helicase. - rhlE, an Escherichia coli putative RNA helicase. - srmB, an Escherichia coli protein that shows RNA-dependent ATPase activity. It probably interacts with 23S ribosomal RNA. - Caenorhabditis elegans hypothetical proteins T26G10.1, ZK512.2 and ZK686.2. - Yeast hypothetical protein

YHR065c. - Yeast hypothetical protein YHR169w. - Fission yeast hypothetical protein SpAC31A2.07c. - Bacillus subtilis hypothetical protein yxiN. All these proteins share a number of conserved sequence motifs. Some of them are specific to this family while others are shared by other ATP-binding proteins or by proteins belonging to the helicases `superfamily' [4,E1]. One of these motifs, called the 'D-E-A-D-box', represents a special version of the B motif of ATP-binding proteins. Some other proteins belong to a subfamily which have His instead of the second Asp and are thus said to be 'D-E-A-H-box' proteins [3,5,6,E1]. Proteins currently known to belong to this subfamily are: - PRP2, PRP16, PRP22 and PRP43. These yeast proteins are all involved in various ATP-requiring steps of the pre-mRNA splicing process. - Fission yeast prh1, which my be involved in pre-mRNA splicing. - Male-less (mle), a Drosophila protein required in males, for dosage compensation of X chromosome linked genes. - RAD3 from yeast. RAD3 is a DNA helicase involved in excision repair of DNA damaged by UV light, bulky adducts or cross-linking agents. Fission yeast rad15 (rhp3) and mammalian DNA excision repair protein XPD (ERCC-2) are the homologs of RAD3. - Yeast CHL1 (or CTF1), which is important for chromosome transmission and normal cell cycle progression in G(2)/M. - Yeast TPS1. - Yeast hypothetical protein YKL078w. - Caenorhabditis elegans hypothetical proteins C06E1.10 and K03H1.2. - Poxviruses' early transcription factor 70 Kd subunit which acts with RNA polymerase to initiate transcription from early gene promoters. - I8, a putative vaccinia virus helicase. - hrpA, an Escherichia coli putative RNA helicase. Signature patterns for both subfamilies were developed.

Consensus pattern: ~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)](2)-D-E-A-D-~~[RKEN]~~[RKEN (SEQ ID NO: 635)]-x-[LIVMFYGSTN

Consensus pattern: ~~[GSAH]~~[GSAH (SEQ ID NO: 198)]-x-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)](3)-D-E-~~[ALIV]~~[ALIV (SEQ ID NO: 8)]-H-~~[NECR]~~[NECR (SEQ ID NO: 564)]

Note: proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop) (see the relevant entry <PDOC00017

[ 1] Schmid S.R., Linder P. Mol. Microbiol. 6:283-292(1992).

[ 2] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).

[ 3] Wassarman D.A., Steitz J.A. Nature 349:463-464(1991).

[ 4] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

[ 5] Harosh I., Deschavanne P. Nucleic Acids Res. 19:6331-6331(1991).

[ 6] Koonin E.V., Senkevich T.G. J. Gen. Virol. 73:989-993(1992).

132. (DHBP_synthase) 3,4-dihydroxy-2-butanone 4-phosphate synthase

3,4-Dihydroxy-2-butanone 4-phosphate is biosynthesized from ribulose 5-phosphate and serves as the biosynthetic precursor for the xylene ring of riboflavin. Sometimes found as a bifunctional enzyme with GTP_cyclohydro2.

Richter G, Krieger C, Volk R, Kis K, Ritz H, Gotze E, Bacher A, Methods Enzymol 1997;280:374-382.

133. (DHDPS) Dihydrodipicolinate synthetase signatures

Dihydrodipicolinate synthetase (EC 4.2.1.52) (DHDPS) [1] catalyzes, in higher plants chloroplast and in many bacteria (gene dapA), the first reaction specific to the biosynthesis of lysine and of diaminopimelate. DHDPS is responsible for the condensation of aspartate semialdehyde and pyruvate by aping-pong mechanism in which pyruvate first binds to the enzyme by forming a Schiff-base with a lysine residue. Three other proteins are structurally related to DHDPS and probably also act via a similar catalytic mechanism: - Escherichia coli N-acetylneuraminate lyase (EC 4.1.3.3) (gene nanA), which catalyzes the condensation of N-acetyl-D-mannosamine and pyruvate to form N-acetylneuraminate. - Rhizobium meliloti protein mosA [3], which is involved in the biosynthesis of the rhizopine 3-o-methyl-scyllo-inosamine. - Escherichia coli hypothetical protein yjhH. Two signature patterns for these enzymes were developed . The first one is centered on highly conserved region in the N-terminal part of these proteins. The second signature contains a lysine residue which has been shown, in Escherichia coli dapA [2], to be the one that forms a Schiff-base with the substrate.

Consensus pattern: [GSA]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x(2)-G-[ST]-[TG]-G-E-~~[GASNF]~~[GASNF (SEQ ID NO: 174)]-x(6)- [EQ] -

Consensus pattern: Y-[DNS]-~~[LIVMFA]~~[LIVMFA (SEQ ID NO: 403)]-P-x(2)-[ST]-x(3)-~~[LIVMG]~~[LIVMG (SEQ ID NO: 486)]-x(13,14)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]- x-[SGA]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-K-~~[DEQAF]~~[DEQAF (SEQ ID NO: 65)]-~~[STAC]~~[STAC (SEQ ID NO: 681)] [K is involved in Schiff-base formation]-

[ 1] Kaneko T., Hashimoto T., Kumpaisal R., Yamada Y. J. Biol. Chem. 265:17451-17455(1990).

[ 2] Laber B., Gomis-Rueth F.-X., Romao M.J., Huber R. Biochem. J. 288:691-695(1992).

[ 3] Murphy P.J., Trenz S.P., Grzemski W., de Bruijn F.J., Schell J. J. Bacteriol. 175:5193-5204 (1993).

134. (DHOdehase) Dihydroorotate dehydrogenase signatures

Dihydroorotate dehydrogenase (EC 1.3.3.1) (DHOdehase) catalyzes the fourth step in the de novo biosynthesis of pyrimidine, the conversion of dihydroorotate into orotate. DHOdehase is a ubiquitous FAD flavoprotein. In bacteria (gene pyrD), DHOdease is located on the inner side of the cytosolic membrane. In some yeasts, such as in Saccharomyces cerevisiae (gene URA1), it is a cytosolic protein while in other eukaryotes it is found in the mitochondria [1]. The sequence of DHOdease is rather well conserved and two signature patterns were developed specific to this enzyme. The first corresponds to a region in the N-terminal section of the enzyme while the second is located in the C-terminal section and seems to be part of the FAD-binding domain.

Consensus pattern[GS]-x(4)-[GK]-~~[GSTA]~~[GSTA (SEQ ID NO: 217)]-~~[LIVFSTA]~~[LIVFSTA (SEQ ID NO: 368)]-[GT]-x(3)-[NQR]-x-G-[NHY]-x(2)-P-[RT]
Consensus pattern~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-[GSA]-x-G-G-[IV]-x-~~[STGDN]~~[STGDN (SEQ ID NO: 746)]-x(3)-[ACV]-x(6)-G-A

[ 1] Nagy M., Lacroute F., Thomas D. Proc. Natl. Acad. Sci. U.S.A. 89:8966-8970(1992).

135. (DMRL_synthase) 6,7-dimethyl-8-ribityllumazine synthase

136. (DNA_methylase) C-5 cytosine-specific DNA methylases signatures

C-5 cytosine-specific DNA methylases (EC 2.1.1.73) (C5 Mtase) are enzymes that specifically methylate the C-5 carbon of cytosines in DNA [1,2,3]. Such enzymes are found in the proteins described below. - As a component of type II restriction-modification systems in prokaryotes and some bacteriophages. Such enzymes recognize a specific DNA sequence

where they methylate a cytosine. In doing so, they protect DNA from cleavage by type II restriction enzymes that recognize the same sequence. The sequences of a large number of type II C-5 Mtases are known. - In vertebrates, there are a number of C-5 Mtases that methylate CpG dinucleotides. The sequence of the mammalian enzyme is known.C-5 Mtases share a number of short conserved regions. Two of them were selected. The first is centered around a conserved Pro-Cys dipeptide in which the cysteine has been shown [4] to be involved in the catalytic mechanism; it appears to form a covalent intermediate with the C6 position of cytosine. The second region is located at the C-terminal extremity in type-II enzymes

Consensus pattern: ~~[DENKS]~~[DENKS (SEQ ID NO: 28)]-x-~~[FLIV]~~[FLIV (SEQ ID NO: 120)]-x(2)-~~[GSTC]~~[GSTC (SEQ ID NO: 240)]-x-P-C-x(2)-~~[FYWLIM]~~[FYWLIM (SEQ ID NO: 151)]-S [C is the active site residue]-

Consensus pattern: ~~[RKQGTF]~~[RKQGTF (SEQ ID NO: 643)]-x(2)-G-N-~~[STAG]~~[STAG (SEQ ID NO: 690)]-~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-x(3)-~~[LIVMT]~~[LIVMT (SEQ ID NO: 518)]-x(3)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]- x(3)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-

[ 1] Posfai J., Bhagwat A.S., Roberts R.J. Gene 74:261-263(1988).

[ 2] Kumar S., Cheng X., Klimasauskas S., Mi S., Posfai J., Roberts R.J., Wilson G.G. Nucleic Acids Res. 22:1-10(1994).

[ 3] Lauster R., Trautner T.A., Noyer-Weidner M. J. Mol. Biol. 206:305-312(1989).

[ 4] Chen L., McMillan A.M., Chang W., Ezak-Nipkay K., Lane W.S., Verdine G.L. Biochemistry 30:11018-11025(1991).

137. (DNAphotolyase) DNA photolyases class 2 signatures

Deoxyribodipyrimidine photolyase (EC 4.1.99.3) (DNA photolyase) [1,2] is a DNArepair enzyme. It binds to UV-damaged DNA containing pyrimidine dimers and, upon absorbing a near-UV photon (300 to 500 nm), breaks the cyclobutane ring joining the two pyrimidines of the dimer. DNA photolyase is an enzyme that requires two choromophore-cofactors for its activity: a reduced FADH2 and either 5,10-methenyltetrahydrofolate (5,10-MTFH) or an oxidized 8-hydroxy-5-deazaflavin (8-HDF) derivative (F420). The folate or deazaflavin chromophore appears to function as an antenna, while the FADH2 chromophore is thought to

be responsible for electron transfer. On the basis of sequence similarities[3] DNA photolyases can be grouped into two classes. The second class contains enzymes from Myxococcus xanthus, methanogenic archaebacteria, insects, fish and marsupial mammals. It is not yet known what second cofactor is bound to class 2 enzymes. There are a number of conserved sequence regions in all known class 2 DNAphotolyases, especially in the C-terminal part. Two of these regions were selected as signature patterns.

Consensus pattern: F-x-E-E-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-R-R-E-L-x(2)-N-F-

Consensus pattern: G-x-H-D-x(2)-W-x-E-R-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-F-G-K-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-R-[FY]-M-N-

[ 1] Sancar G.B., Sancar A. Trends Biochem. Sci. 12:259-261(1987).

[ 2] Jorns M.S. Biofactors 2:207-211(1990).

[ 3] Yasui A., Eker A.P.M., Yasuhira S., Yajima H., Kobayashi T., Takao M., Oikawa A. EMBO J. 13:6143-6151(1994).

(DNAphotolyase2) DNA photolyases class 1 signatures

Deoxyribodipyrimidine photolyase (EC 4.1.99.3) (DNA photolyase) [1,2] is a DNA repair enzyme. It binds to UV-damaged DNA containing pyrimidine dimers and ,upon absorbing a near-UV photon (300 to 500 nm), breaks the cyclobutane ring joining the two pyrimidines of the dimer. DNA photolyase is an enzyme that requires two choromophore-cofactors for its activity: a reduced FADH2 and either 5,10-methenyltetrahydrofolate (5,10-MTFH) or an oxidized 8-hydroxy-5-deazaflavin (8-HDF) derivative (F420). The folate or deazaflavin chromophore appears to function as an antenna, while the FADH2 chromophore is thought to be responsible for electron transfer. On the basis of sequence similarities[3] DNA photolyases can be grouped into two classes. The first class contains enzymes from Gram-negative and Gram-positive bacteria, the halophilic archaebacteria Halobacterium halobium, fungi and plants. Class 1 enzymes bind either 5,10-MTHF (E.coli, fungi, etc.) or 8-HDF (S.griseus, H.halobium).This family also includes Arabidopsis cryptochromes 1 (CRY1) and 2 (CRY2),which are blue light photoreceptors that mediate blue light-induced gene expression. There are a number of conserved sequence regions in all known class 1 DNA photolyases, especially in the C-terminal part. Two of these regions were selected as signature patterns

Consensus pattern: T-G-x-P-~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-D-A-x-M-[RA]-x-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-

Consensus pattern: [DN]-R-x-R-~~[LIVM]~~[LIVM (SEQ ID NO: 382)](2)-x-[STA](2)-F-~~[LIVMFA]~~[LIVMFA (SEQ ID NO: 403)]-x-K-x-L-x(2,3)- W-[KRQ]-

[ 1] Sancar G.B., Sancar A. Trends Biochem. Sci. 12:259-261(1987).

[ 2] Jorns M.S. Biofactors 2:207-211(1990).

[ 3] Yasui A., Eker A.P.M., Yasuhira S., Yajima H., Kobayashi T., Takao M., Oikawa A. EMBO J. 13:6143-6151(1994).

[ 4] Lin C., Ahmad M., Cashmore A.R. Plant J. 10:893-902(1996).

138. (DNA_pol_A)

DNA polymerase family A signature

Replicative DNA polymerases (EC 2.7.7.7) are the key enzymes catalyzing the accurate replication of DNA. They require either a small RNA molecule or a protein as a primer for the de novo synthesis of a DNA chain. On the basis of sequence similarities a number of DNA polymerases have been grouped together [1,2,3] under the designation of DNA polymerase family A. The polymerases that belong to this family are listed below.

- Escherichia coli and various other bacterial polymerase I (gene polA).
- Thermus aquaticus Taq polymerase.
- Bacteriophage sp01 polymerase.
- Bacteriophage sp02 polymerase.
- Bacteriophage T5 polymerase.
- Bacteriophage T7 polymerase.
- Mycobacteriophage L5 polymerase.
- Yeast mitochondrial polymerase gamma (gene MIP1).

Five regions of similarity are found in all the above polymerases. One of these conserved regions, known as 'motif B' [1], is located in a domain which, in Escherichia coli polA, has been shown to bind deoxynucleotide triphosphate substrates; it contains a conserved tyrosine which has been shown, by photo- affinity labelling, to be in the active site; a conserved

lysine, also part of this motif, can be chemically labelled, using pyridoxal phosphate. This conserved region was used as a signature for this family of DNA polymerases.

Consensus patternR-x(2)-~~[GSAV]~~[GSAV (SEQ ID NO: 208)]-K-x(3)-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-[AGQ]-x(2)-Y-x(2)-[GS]-x(3)- ~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)]
Sequences known to belong to this class detected by the pattern ALL.

[ 1] Delarue M., Poch O., Todro N., Moras D., Argos P. Protein Eng. 3:461-467(1990).
[ 2] Ito J., Braithwaite D.K. Nucleic Acids Res. 19:4045-4057(1991).
[ 3] Braithwaite D.K., Ito J. Nucleic Acids Res. 21:787-802(1993).

139. DNA_pol_viral_C
DNA polymerase (viral) C-terminal domain
Number of members: 128

140. (DNA_topoisoII)
DNA topoisomerase II signature
DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break. Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes gyrA and gyrB [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes parC and parE). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

<----------About-1400-residues---------->

[----------Protein 39-*-----][----Protein 52----] Phage T4

[----------gyrB-------*-----][--------gyrA----------] Prokaryote II

Archaebacteria

[----------parE-------*-----][--------parD----------] Prokaryote IV

[--------------------*----------] Eukaryote and

ASF

'*': Position of the pattern.

As a signature pattern for this family of proteins, a region that contains a highly conserved pentapeptide was selected. The pattern is located in gyrB, in parE, and in protein 39 of phage T4 topoisomerase.

Consensus pattern~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)]-x-E-G-[DN]-S-A-x-~~[STAG]~~[STAG (SEQ ID NO: 690)] Sequences known to belong to this class detected by the pattern ALL.

[ 1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).

[ 2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).

[ 3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).

[ 4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

141. (DSPc) Tyrosine specific protein phosphatases signature and profiles

Tyrosine specific protein phosphatases (EC 3.1.3.48) (PTPase) [1 to 5] are enzymes that catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation. Multiple forms of PTPase have been characterized and can be classified into two categories: soluble PTPases and transmembrane receptor proteins that contain PTPase domain(s). The currently known PTPases are listed below: Soluble PTPases. - PTPN1 (PTP-1B). - PTPN2 (T-cell PTPase; TC-PTP). - PTPN3 (H1) and PTPN4 (MEG), enzymes that contain an N-terminal band 4.1- like domain (see <PDOC00566>) and could act at junctions between the membrane and cytoskeleton. - PTPN5 (STEP). - PTPN6 (PTP-1C; HCP; SHP) and PTPN11

(PTP-2C; SH-PTP3; Syp), enzymes which contain two copies of the SH2 domain at its N-terminal extremity. The Drosophila protein corkscrew (gene csw) also belongs to this subgroup. - PTPN7 (LC-PTP; Hematopoietic protein-tyrosine phosphatase; HePTP). - PTPN8 (70Z-PEP). - PTPN9 (MEG2). - PTPN12 (PTP-G1; PTP-P19). - Yeast PTP1. - Yeast PTP2 which may be involved in the ubiquitin-mediated protein degradation pathway. - Fission yeast pyp1 and pyp2 which play a role in inhibiting the onset of mitosis. - Fission yeast pyp3 which contributes to the dephosphorylation of cdc2. - Yeast CDC14 which may be involved in chromosome segregation. - Yersinia virulence plasmid PTPAses (gene yopH). - Autographa californica nuclear polyhedrosis virus 19 Kd PTPase.Dual specificity PTPases. - DUSP1 (PTPN10; MAP kinase phosphatase-1; MKP-1); which dephosphorylates MAP kinase on both Thr-183 and Tyr-185. - DUSP2 (PAC-1), a nuclear enzyme that dephosphorylates MAP kinases ERK1 and ERK2 on both Thr and Tyr residues. - DUSP3 (VHR). - DUSP4 (HVH2). - DUSP5 (HVH3). - DUSP6 (Pyst1; MKP-3). - DUSP7 (Pyst2; MKP-X). - Yeast MSG5, a PTPase that dephosphorylates MAP kinase FUS3. - Yeast YVH1. - Vaccinia virus H1 PTPase; a dual specificity phosphatase. Receptor PTPases. Structurally, all known receptor PTPases, are made up of a variable length extracellular domain, followed by a transmembrane region and a C-terminalcatalytic cytoplasmic domain. Some of the receptor PTPases contain fibronectintype III (FN-III) repeats, immunoglobulin-like domains, MAM domains orcarbonic anhydrase-like domains in their extracellular region. The cytoplasmic region generally contains two copies of the PTPAse domain. The first seems to have enzymatic activity, while the second is inactive but seems to affect substrate specificity of the first. In these domains, the catalytic cysteine is generally conserved but some other, presumably important, residues are not. In the following table, the domain structure of known receptor PTPases is shown: Extracellular Intracellular ---------- ---------- Ig FN-3 CAH MAM PTPaseLeukocyte common antigen (LCA) (CD45) 0 2 0 0 2Leukocyte antigen related (LAR) 3 8 0 0 2 Drosophila DLAR 3 9 0 0 2Drosophila DPTP 2 2 0 0 2PTP-alpha (LRP) 0 0 0 0 2PTP-beta 0 16 0 0 1PTP-gamma 0 1 1 0 2PTP-delta 0 >7 0 0 2 PTP-epsilon 0 0 0 0 2PTP-kappa 1 4 0 1 2PTP-mu 1 4 0 1 2PTP-zeta 0 1 1 0 2PTPase domains consist of about 300 amino acids. There are two conserved cysteines, the second one has been shown to be absolutely required for activity. Furthermore, a number of conserved residues in its immediate vicinity have also been shown to be important. A signature pattern for PTPase domains was derived centered on the active site cysteine. There are three profiles for PTPases, the first one spans the complete domain and is not specific to any subtype. The second profile is specific to dual-specificity PTPases and the third one to the PTP subfamily

Consensus pattern: ~~[LIVMF]~~[LIVMF (SEQ ID NO: 402)]-H-C-x(2)-G-x(3)-[STC]-~~[STAGP]~~[STAGP (SEQ ID NO: 707)]-x-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)] [C is the active site residue]-

[ 1] Fischer E.H., Charbonneau H., Tonks N.K. Science 253:401-406(1991).

[ 2] Charbonneau H., Tonks N.K. Annu. Rev. Cell Biol. 8:463-493(1992).

[ 3] Trowbridge I.S. J. Biol. Chem. 266:23517-23520(1991).

[ 4] Tonks N.K., Charbonneau H. Trends Biochem. Sci. 14:497-500(1989).

[ 5] Hunter T. Cell 58:1013-1016(1989).

142. (DUF10) Uncharacterized protein family UPF0076 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Goat antigen UK114, a human homolog and the rat corresponding protein which is known as perchloric acid soluble protein (PSP1). PSP1 [2] may inhibit an initiation stage of cell-free protein synthesis. - Mouse heat-responsive protein HRSP12. - Yeast chromosome V hypothetical protein YER057c. - Yeast chromosome IX hypothetical protein YIL051c. - Caenorhabditis elegans hypothetical protein C23G10.2. - Escherichia coli hypothetical protein ycdK. - Escherichia coli hypothetical protein yhaR. - Escherichia coli hypothetical protein yjgF and HI0719, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yoaB. - Bacillus subtilis hypothetical protein yabJ. - Haemophilus influenzae hypothetical protein HI1627. - Helicobacter pylori hypothetical protein HP0944. - Lactococcus lactis aldR. - Myxococcus xanthus dfrA. - Synechocystis strain PCC 6803 hypothetical protein slr0709. - Rhizobium strain NGR234 symbiotic plasmid hypothetical protein y4sK. - Pyrococcus horikoshii hypothetical protein PH0854.These are small proteins of around 15 Kd whose sequence is highly conserved.As a signature pattern, a well conserved region located in the C-terminal part of these proteins was selected.

Consensus pattern: [PA]-~~[ASTPV]~~[ASTPV (SEQ ID NO: 12)]-R-~~[SACVF]~~[SACVF (SEQ ID NO: 654)]-x-~~[LIVMFY]~~[LIVMFY (SEQ ID NO: 434)]-x(2)-~~[GSAKR]~~[GSAKR (SEQ ID NO: 202)]-x-~~[LMVA]~~[LMVA (SEQ ID NO: 546)]- x(5,8)-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]-E-[MI]-

[ 1] Bairoch A. Unpublished observations (1995).

[ 2] Oka T., Tsuji H., Noda C., Sakai K., Hong Y.-M., Suzuki I., Munoz S., Natori Y. J. Biol. Chem. 270:30060-30067(1995).

143. (DUF3)Domain of Unknown Function 3

Domain apparently occurring exclusively in eubacteria. Unknown function.

144. (DUF6) Integral membrane protein

This family includes many hypothetical membrane proteins of unknown function. Many of the proteins contain two copies of the aligned region.

145. (DUF7) Integral membrane protein

This family includes many hypothetical membrane proteins of unknown function. Swiss:P14502 has been implicated in resistance to ethidium bromide.

146. (DapB) Dihydrodipicolinate reductase signature

Dihydrodipicolinate reductase (EC 1.3.1.26) catalyzes the second step in the biosynthesis of diaminopimelic acid and lysine, the NAD or NADP-dependent reduction of 2,3-dihydrodipicolinate into 2,3,4,5-tetrahydrodipicolinate. This enzyme is present in bacteria (gene dapB) and higher plants. As a signature pattern the best conserved region in this enzyme was selected. It is located in the central section and is part of the substrate-binding region [1].

Consensus pattern: E-[IV]-x-E-x-H-x(3)-K-x-D-x-P-S-G-T-A-

[ 1] Scapin G., Blanchard J.S., Sacchettini J.C. Biochemistry 34:3502-3512(1995).

147. DedA family

This family combines the DedA related proteins and YIAN/YGIK family. Members of this family are not functionally characterised. These proteins contain multiple predicted transmembrane regions.

148. DegT/DnrJ/EryC1/StrS family

The members of this family exhibit some characteristics of the sensor protein of two-component signal transduction systems, however none of the members show any sequence similarity to these protein kinases. The members of this family do have the typical helix-turn-helix motif of DNA binding proteins.

[1] Stutzman-Engwall KJ, Otten SL, Hutchinson CR, J Bacteriol 1992;174:144-154.

149. (Desaturase) Fatty acid desaturases signatures

Fatty acid desaturases (EC 1.14.99.-) are enzymes that catalyze the insertion of a double bond at the delta position of fatty acids. There seems to be two distinct families of fatty acid desaturases which do not seem to be evolutionary related. Family 1 is composed of: - Stearoyl-CoA desaturase (SCD) (EC 1.14.99.5) [1]. SCD is a key regulatory enzyme of unsaturated fatty acid biosynthesis. SCD introduces a cis double bond at the delta(9) position of fatty acyl-CoA's such as palmitoleoyl- and oleoyl-CoA. SCD is a membrane-bound enzyme that is thought to function as a part of a multienzyme complex in the endoplasmic reticulum of vertebrates and fungi. As a signature pattern for this family a conserved region in the C-terminal part of these enzymes was selected, this region is rich in histidine residues and in aromatic residues. Family 2 is composed of: - Plants stearoyl-acyl-carrier-protein desaturase (EC 1.14.99.6) [2], these enzymes catalyze the introduction of a double bond at the delta(9) position of steraoyl-ACP to produce oleoyl-ACP. This enzyme is responsible for the conversion of saturated fatty acids to unsaturated fatty acids in the synthesis of vegetable oils. - Cyanobacteria desA [3] an enzyme that can introduce a second cis double bond at the delta(12) position of fatty acid bound to membranes glycerolipids. DesA is involved in chilling tolerance; the phase transition temperature of lipids of cellular membranes being dependent on the degree of unsaturation of fatty acids of the membrane lipids. As a signature pattern for this family a conserved region in the C-terminal part of these enzymes was selected.

Consensus pattern: G-E-x-[FY]-H-N-[FY]-H-H-x-F-P-x-D-Y-

Consensus pattern: [ST]-[SA]-x(3)-[QR]-[LI]-x(5,6)-D-Y-x(2)-~~[LIVMFYW]~~[LIVMFYW (SEQ ID NO: 463)]-~~[LIVM]~~[LIVM (SEQ ID NO: 382)]- [DE]-

[ 1] Kaestner K.H., Ntambi J.M., Kelly T.J. Jr., Lane M.D. J. Biol. Chem. 264:14755-14761(1989).

[ 2] Shanklin J., Somerville C.R. Proc. Natl. Acad. Sci. U.S.A. 88:2510-2514(1991).

[ 3] Wada H., Gombos Z., Murata N. Nature 347:200-203(1990).

150. Dihydroorotase signatures

Dihydroorotase (EC 3.5.2.3) (DHOase) catalyzes the third step in the de novo biosynthesis of pyrimidine, the conversion of ureidosuccinic acid (N-carbamoyl-L-aspartate) into dihydroorotate. Dihydroorotase binds a zinc ion which is required for its catalytic activity [1]. In bacteria, DHOase is a dimer of identical chains of about 400 amino-acid residues (gene pyrC). In higher eukaryotes, DHOase is part of a large multi-functional protein known as 'rudimentary' in Drosophila and CAD in mammals and which catalyzes the first three steps of pyrimidine biosynthesis [2]. The DHOase domain is located in the central part of this polyprotein. In yeasts, DHOase is encoded by a monofunctional protein (gene URA4). However, a defective DHOase domain [3] is found in a multifunctional protein (gene URA2)that catalyzes the first two steps of pyrimidine biosynthesis. The comparison of DHOase sequences from various sources shows [4] that there are two highly conserved regions. The first located in the N-terminal extremity contains two histidine residues suggested [3] to be involved in binding the zinc ion. The second is found in the C-terminal part. Signature patterns for both regions have been developed. Allantoinase (EC 3.5.2.5) is the enzyme that hydrolyzes allantoin intoallantoate. In yeast (gene DAL1) [5], it is the first enzyme in the allanto indegradation pathway; in amphibians [6] and fish it catalyzes the second step in the degradation of uric acid. The sequence of allantoinase is evolutionary related to that of DHOases.

Consensus pattern: D-~~[LIVMFYWSAP]~~[LIVMFYWSAP (SEQ ID NO: 481)]-H-~~[LIVA]~~[LIVA (SEQ ID NO: 351)]-H-~~[LIVF]~~[LIVF (SEQ ID NO: 360)]-[RN]-x-~~[PGANF]~~[PGANF (SEQ ID NO: 598)] [The two H's are probable zinc ligands]-

Consensus pattern: [GA]-[ST]-D-x-A-P-H-x(4)-K-

[ 1] Brown D.C., Collins K.D. J. Biol. Chem. 266:1597-1604(1991).

[ 2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-164(1993).

[ 3] Souciet J.-L., Nagy M., Le Gouar M., Lacroute F., Potier S. Gene 79:59-70(1989).

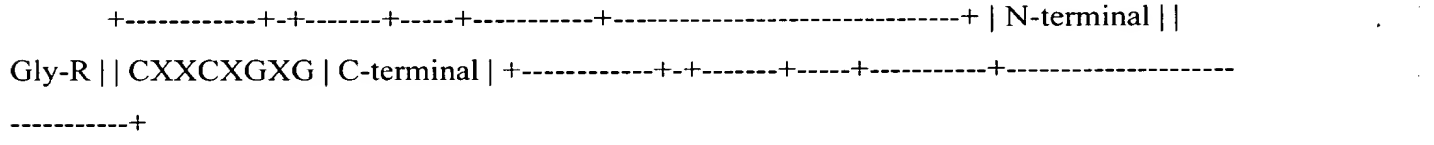
[ 4] Guyonvarch A., Nguyen-Juilleret M., Hubert J.-C., Lacroute F. Mol. Gen. Genet. 212:134-141(1988).

[ 5] Buckholz R.G., Cooper T.G. Yeast 7:913-923(1991).

[ 6] Hayashi S., Jain S., Chu R., Alvares K., Xu B., Erfurth F., Usuda N., Rao M.S., Reddy S.K., Noguchi T., Reddy J.K., Yeldandi A.Y. J. Biol. Chem. 269:12269-12276(1994).

151. dnaJ domains signatures and profile

The prokaryotic heat shock protein dnaJ interacts with the chaperone hsp70-like dnaK protein [1]. Structurally, the dnaJ protein consists of an N- terminal conserved domain (called 'J' domain) of about 70 amino acids, a glycine-rich region ('G' domain') of about 30 residues, a central domain containing four repeats of a CXXCXGXG motif ('CRR' domain) and a C-terminal region of 120 to 170 residues. Such a structure is shown in the following schematic representation:




It has been shown [2] that the 'J' domain as well as the 'CRR' domain are also found in other prokaryotic and eukaryotic proteins which are listed below.

a) Proteins containing both a 'J' and a 'CRR' domain:

- Yeast protein MAS5/YDJ1 which seems to be involved in mitochondrial protein import.
- Yeast protein MDJ1, involved in mitochondrial biogenesis and protein folding.
- Yeast protein SCJ1, involved in protein sorting.
- Yeast protein XDJ1.
- Plants dnaJ homologs (from leek and cucumber).
- Human HDJ2, a dnaJ homolog of unknown function.
- Yeast hypothetical protein YNL077w.

b) Proteins containing a 'J' domain without a 'CRR' domain:

- Rhizobium fredii nolC, a protein involved in cultivar-specific nodulation of soybean.
- Escherichia coli cbpA [3], a protein that binds curved DNA.
- Yeast protein SEC63/NPL1, important for protein assembly into the endoplasmic reticulum and the nucleus.
- Yeast protein SIS1, required for nuclear migration during mitosis.
- Yeast protein CAJ1.
- Yeast hypothetical protein YFR041c.
- Yeast hypothetical protein YIR004w.
- Yeast hypothetical protein YJL162c.
- Plasmodium falciparum ring-infected erythrocyte surface antigen (RESA). RESA, whose function is not known, is associated with the membrane skeleton of newly invaded erythrocytes.
- Human HDJ1.
- Human HSJ1, a neuronal protein.
- Drosophila cysteine-string protein (csp).

A signature pattern for the 'J' domain was developed, based on conserved positions in the C-terminal half of this domain. A pattern for the 'CRR' domain, based on the first two copies of that motif was also developed. A profile for the 'J' domain was also developed.

Consensus pattern: [FY]-x(2)-~~[LIVMA]~~[LIVMA (SEQ ID NO: 383)]-x(3)-~~[FYWHNT]~~[FYWHNT (SEQ ID NO: 146)]-~~[DENQSA]~~[DENQSA (SEQ ID NO: 48)]-x-L-x-[DN]-x(3)- [KR]-x(2)-[FYI]-

Consensus pattern: C-~~[DEGSTHKR]~~[DEGSTHKR (SEQ ID NO: 23)]-x-C-x-G-x-[GK]-~~[AGSDM]~~[AGSDM (SEQ ID NO: 6)]-x(2)-~~[GSNKR]~~[GSNKR (SEQ ID NO: 216)]-x(4,6)-C- x(2,3)-C-x-G-x-G-

[1] Cyr D.M., Langer T., Douglas M.G. Trends Biochem. Sci. 19:176-181(1994).

[2] Bork P., Sander C., Valencia A., Bukau B. Trends Biochem. Sci. 17:129-129(1992).

[3] Ueguchi C., Kaneda M., Yamada H., Mizuno T. Proc. Natl. Acad. Sci. U.S.A. 91:1054-1058(1994).

153. Dwarfin

This family known as the dwarfins also includes the drosophila protein MAD. The N-terminus of MAD can bind to DNA [2].

[1] Yingling JM, Das P, Savage C, Zhang M, Padgett RW, Wang XF, Proc Natl Acad Sci U S A 1996;93:8940-8944. [2] Kim J, Johnson K, Chen HJ, Carroll S, Laughon A, Nature 1997;388:304-308.

154. Dynein light chain type 1 signature

Dynein is a multisubunit microtubule-dependent motor enzyme that acts as the force generating protein of eukaryotic cilia and flagella. The cytoplasmic isoform of dynein acts as a motor for the intracellular retrograde motility of vesicles and organelles along microtubules. Dynein is composed of a number of ATP-binding large subunits, intermediate size subunits and small subunits. Among the small subunits, there is a family [1,2] of highly conserved proteins which consist of: - Chlamydomonas reinhardtii flagellar outer arm dynein 8 Kd and 11 Kd light chains. - Higher eukaryotes cytoplasmic dynein light chain 1. - Yeast cytoplasmic dynein light chain 1 (gene DYN2 or SLC1). - Caenorhabditis elegans hypothetical dynein light chains M18.2 and T26A5.9.These proteins are have from 89 to 120 amino acids. As a signature pattern, A highly conserved region was selected.

Consensus pattern: H-x-I-x-G-[KR]-x-F-[GA]-S-x-V-[ST]-[HY]-E -

[ 1] King S.M., Patel-King R.S. J. Biol. Chem. 270:11445-11452(1995).

[2] Dick T., Ray K., Salz H.K., Chia W. Mol. Cell. Biol. 16:1966-1977(1996).

155. dUTPase

dUTPase hydrolyzes dUTP to dUMP and pyrophosphate.

[1] Cedergren-Zeppezauer ES, Larsson G, Nyman PO, Dauter Z, Wilson KS, Nature 1992;355:740-743. [2] Mol CD, Harris JM, McIntosh EM, Tainer JA, Structure 1996;4:1077-1092.